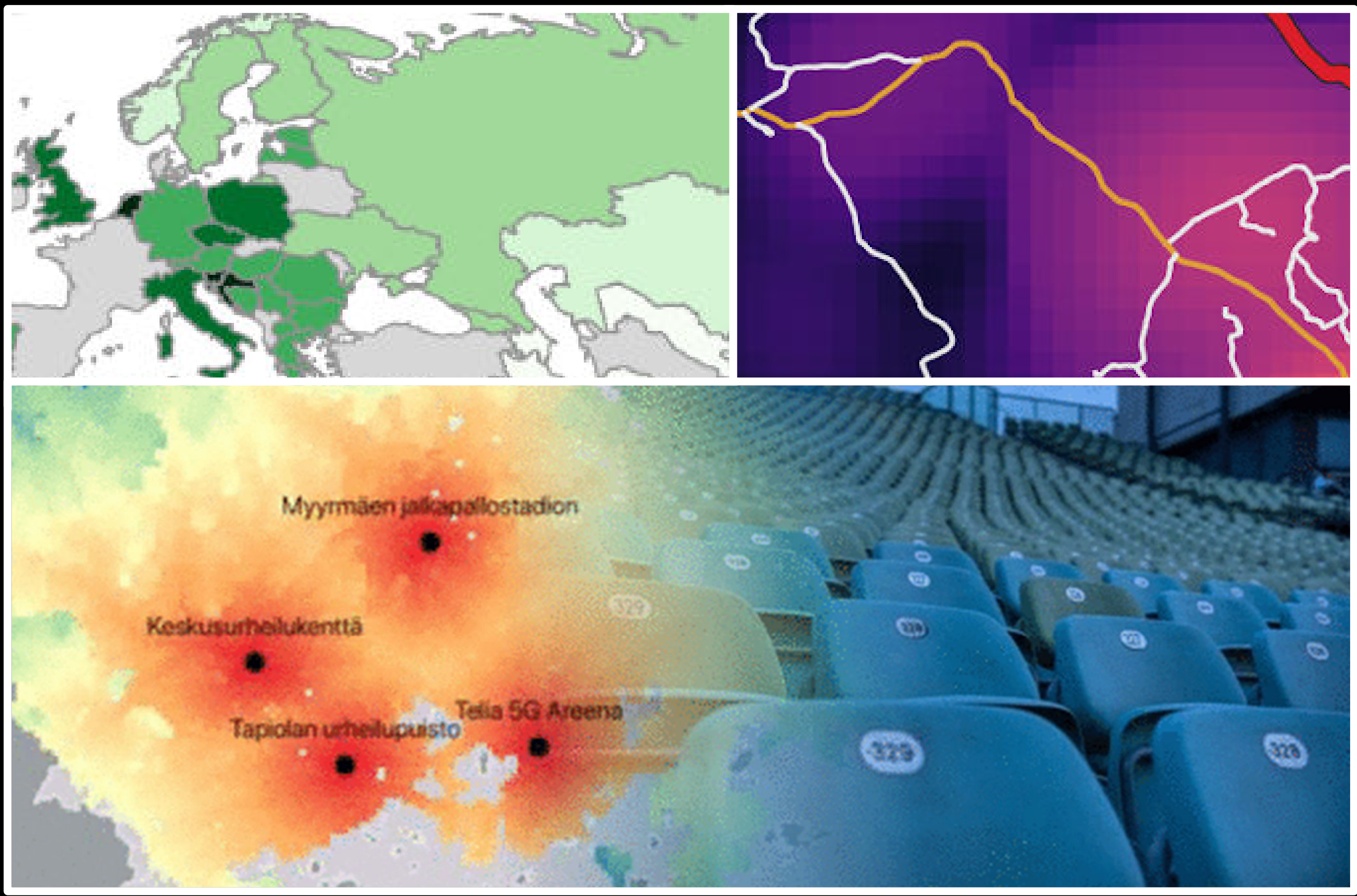
2020 DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY C19



DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY C19

MUUKKONEN, P. (Ed.)

Examples and progress in geodata science

Final report of MSc course at the Department of Geosciences and Geography, University of Helsinki, spring 2020

MUUKKONEN, P. (Ed.)

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

Examples and progress in geodata science:

Final report of MSc course at the Department of Geosciences and Geography, University of Helsinki, spring 2020

EDITOR:
PETTERI MUUKKONEN

DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY C19 / HELSINKI 2020

Publisher:
Department of Geosciences and Geography
Faculty of Science
P.O. Box 64, 00014 University of Helsinki, Finland

Journal:
Department of Geosciences and Geography C19
ISSN-L 1798-7938
ISBN 978-951-51-4938-1 (PDF)
http://helda.helsinki.fi/

Helsinki 2020

Muukkonen, P. (Ed.): Examples and progress in geodata science. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Table of contents

Editor's preface

Examples and progress in geodata science

Geodata science (or geographical data science) has raised its interest and importance during the past years. This is due the more diverse sources of the geographical information and data. Nowadays we can get massive amounts of data for example from the social media. In addition, computation power, technology, data storages, and even cloud-computing have improved a lot. All these improvements and changes have influenced the thinking how different sectors in the society and academia put efforts to gather data, analyse it and deliver outputs and results to various stakeholders. It is not an old c cliché that we are living in the era of information breakout.

In this breakout, there are also room for spatial thinking. I see that geographers have a lot to provide for the different actors using and demanding huge amount of digital data (~ big data). In quite many cases the data includes some sort of spatial element. Not always directly coordinates, but the data may consist of information that can be linked to locations by some sort of georeferencing. We can make georeferencing with the help of addresses, hashtags, other place names, image analyses, joining, database relations and so on. Our thinking of the geographical data (or geodata) has been broadened a lot. And in the future, we must think out-of-the-box more than before. We can't even imagine right now what kind of digital geographical data sources we might have in the future. But we know already that data amounts are large (~ big data) and data consists of both "traditional" research data as well as voluntary data or data for example from the social media. With social media data researcher are studying user groups' behaviour, values, or movement, but the data is bot originally designed for the research purposes. So, the definition of the big data is discordant – some are defining big data as a huge amount of any data, but some are defining it as a data that is not originally purposed purely for research purposes (such as social media data, data from video games, data from mobile apps such to mention few). One can consider that for example voluntary data of bird or plant species observations belongs also under the definition for big data. Therefore, I repeat again that in the future the variety of geographical data sources will become more diverse.

There is a growing scientific and societal need for digital geographical data sciences. Therefore, our duty in the university is to educate future's workers, scientists, and specialists to work with more diverse digital geographical data and larger data amounts. This publication is an output of university level course in geography "*GEOG-G303 GIS Project Work*". In this course, group of students worked together with our department's researchers and teachers to do practical, topical, and real projects assignments. Researchers worked as mentors and clients during the project course. In this way all project works were directly linked to actual research work and research projects. This built a strong link between the teaching and the research, which are the main two focus and goals in the university. In addition, in this way students got and idea and a direct contact surface what are the current advances and outcomes in the geographical data sciences.

This publication consists of seven chapter covering all those project assignments. Common for all the chapters and project assignments is that they all deal with digital geodata. The Chapter I how to use Twitter statistics when tracking patterns of tourism. Incidentally, the Chapter VII uses also social media data. The chapter VII show a demonstration how to utilize Flickr social media data when studying wildlife conservation and conservation related tourism. There occur also other types of movement than tourism. Chapter VI shows demonstrations how geographical data science can be used when studying actual movement of people inside the metropolitan. And what kind of challenges one might face while processing geographical digital data. In addition, Chapter III shows a workflow and a case study how to combine various data type and data sources when analysing reachability (~ accessibility) and travel time pattern of a common service network – in this case official sports facilities. In geographical data sciences it is typical that one combines several databases and different data types together. And those processes should be automized and documented properly. In this course, geography students got valuable knowledge how to plan, execute, and document projects related on geographical data sciences.

Chapters II, IV, and V show how diverse the current field of geographical data sciences really is. The Chapter II review and discuss how geographical data sciences can help with current pandemic started globally in spring 2020. The pandemic hit the Finland exactly at the middle of this course. This challenged our students and teachers because the campus and the university were locked with only a short notice. Luckily, we managed to finalize this course with flexibility. Flexibility was needed also with the Chapter IV, in which we wanted to do some data converting for our VR utilities. The pandemic lock-down forced us to skip all actual data converting and work-flow documentation, which was our preliminary goal. Now the Chapter IV is a literature review about 3D data and virtual reality (VR). VR has broken through in the field of geographical data sciences and GIS. In our department we have two brand new VR utilities (VR cabins).

All the chapters in this publication demonstrate the wide variety of need for geographical data sciences. The Chapter V shows that geographical data science skills and methodology is needed also in the biogeographical research. So, geographical data sciences and its methods and approaches are needed in the both ends of the spectrum: in the "hard and cold" quantitative physical geography as well as in "soft" qualitative data. The common thing for all this variety in geographical research is that digital data is almost every time present, data amounts are growing, data types and sources have more and more variation, and a good documentation is needed.

This publication continues the series of course publications from the course "GEOG-G303 GIS Project Work":

Kujala, S. & Muukkonen, P. (Eds.) (2019). *GIS applications in teaching and research*. Department of Geosciences and Geography C17. https://helda.helsinki.fi/handle/10138/309007

Tyystjärvi, V. & Muukkonen, P. (Eds.) (2018). *Creating, managing, and analysing geospatial data and databases in geographical themes*. Department of Geosciences and Geography C14. https://helda.helsinki.fi/handle/10138/254913

Editor
Petteri Muukkonen
University of Helsinki

Aagesen, H., Levin, A., Ojansuu, S., Redding, A., Muukkonen, P. & Järv, O. (2020). Using Twitter data to evaluate tourism in Finland – A comparison with official statistics. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 3–16. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter I

Using Twitter data to evaluate tourism in Finland – A comparison with official statistics

Aagesen H., Levlin, A., Ojansuu, S., Redding A., Muukkonen, P. & Järv, O.*

havard.aagesen@helsinki.fi, *University of Helsinki*
anna.levlin@helsinki.fi, *University of Helsinki*
sirpa.ojansuu@helsinki.fi, *University of Helsinki*
alisa.redding@helsinki.fi, *University of Helsinki*
petteri.muukkonen@helsinki.fi, *University of Helsinki*
olle.jarv@helsinki.fi, *University of Helsinki*

* *Corresponding author* olle.jarv@helsinki.fi

Abstract

The aim for this study was to determine the applicability of Twitter data as a reliable source for studying and documenting human movement. Increasingly, social media data can be used to study many facets of human geography, movement, residence, places of interest, activities, etc. Twitter is not one of the most widely used social media platforms, nor is it one where geo-tagging is very popular. It is reported that only 0.85% of all Tweets are geotagged (Sloan and Morgan 2015). Nevertheless, the possibilities to access human movement even within that small percentage are strong. The accessibility of data through voluntary location sharing is still relatively new, so, there are always new opportunities emerging to learn about human patterns. This study uses Twitter data geotagged in Finland from October 2017 – October 2019. That data is then compared to *Visit Finland's* Statistics Service Rudolf with data from *Statistics Finland* to determine how well Twitter data performs as a substitute to more traditional sources of data. Additionally, Finnish user data is analysed to know how far the average Finnish person travels outside home, and the most popular regions to visit according to unique user tweets.

Keywords: Big data; Geotagging; Tourism; Twitter

Introduction

Social media is a significant part of the constantly created and stored big data around the world. This enormous amount of social media data provides researchers new perspectives and outlets for studying human patterns and behaviour (Hawelka et al. 2014; Toivonen et al. 2019). Developments in these kinds of data platforms gives researchers' access to user-created content in both real-time and archived data. It allows us to connect users to geographical locations around the world, analysing spatial and temporal patterns in demographics, movements, consumption habits, use of urban greens and beyond (Heikinheimo et al. 2020; Toivonen et al. 2019). Certainly, there are many challenges with using social media data regarding the representativeness of different group of people and the unequal distribution social media activity over space and time, for example. Given challenges stem from the fact that social media use is a voluntary activity that not everyone chooses to use. Thus, social media is not an absolute truth and its content is entirely in the control of the user (Miller and Goodchild 2014). However, the benefits of user-generated data open multiple possibilities to uncover societal processes and phenomena that is not feasible to examine with traditional data sources such as register, survey and interview data.

For instance, tourism can be challenging to gather statistics on, especially to know who the visitors are and how they cross country borders and move within a destination. Yet, this is crucial information as tourism serves an important role in many countries' economies and makes up a large portion of global human movement (Gheasi et al. 2011). In the EU, local and national governments keep track of tourism, predominantly counting overnight stays in hotels. While border crossings are a popular form of tracking tourism statistics in many other countries, "the Schengen agreement – improving free movement of persons by abolishing border controls between countries in the Schengen area – has affected the reliability and feasibility from a methodological and financial point of view, of conducting border surveys" (Eurostat 2014). Additionally, border crossings cannot provide very detailed demographic information due to anonymity requirements of passenger information, which would in many cases prevent the knowledge on whether a passenger's flight was an arrival at final destination or a layover. Similarly, accessing such information requires "ongoing and permanent co-operation between the bodies responsible for generating the files," such as police and immigration controls, increasing the overall difficulty of gathering complete tourism statistics (International Air Transport Association 2002). At the same time, along with the globalization, airlines offer more travel destinations and competition for the best prices increases amongst airlines. The overall travel increases worldwide, and regional economies are constantly looking for ways to improve their visitor monitoring systems and understanding of visitation preferences.

One potential solution to enhance tourism information could the use of social media data. Over the past decade, the proliferation of smart phones and sensor-networks has allowed for the substantial increase in large geo-tagged datasets (Chua et. al 2016; Toivonen et al. 2019). In the global digital age, leisure and tourism activities are well shared via social media platforms (Tenkanen et al. 2017). Travelling can be even dictated by social media as social media users get travel recommendations from each other and similarly, document their journeys for their friends to see. The information stored in social media posts provide the ideal datasets for understanding both travel trends (location and distance) and travel preferences (Hawelka et al. 2014; Tenkanen et al. 2017). For example, Hausmann et al. (2018) used social media as a tool for studying preferences for nature-based experiences in the context of ecotourism. They also suggest that protected nature areas can generate more political support for their continued protection if there is statistical evidence of the popularity and preferences of tourists for these kinds of spaces, and for the biodiversity found there.

Statistical evidence is most easily found through social media due to the public accessibility and availability of the data (Hawelka et al. 2014). Geotagged posts allow to reveal movements of tourists in way that is not possible with surveys. This makes it possible to discover not only new points of interest, but the movements visitors make within a destination country. According to Toivonen et al. (2019), the value of user-generated content lies also in the metadata embedded in photos or texts, which includes who, where and when posts or photos were made, and what activities people are sharing. This information allows analysis of tourist flows and volumes at various geographical scales and reasons behind visiting destination locations. However, social media data has not been used in tourism statistics to date because of the biases regarding user profile of social media platforms and the voluntary generation of location-based social media, and the small sample size (Saluveer et al., 2020). This raises the question of providing representative tourism statistics.

Nevertheless, Twitter data has the potential for tourism statistics (Hawelka et al 2014; Tenkanen et al. 2017). Twitter ranks only 13th out of the 15 most used social networking platforms worldwide with roughly 340 million unique users (Kemp 2020). Additionally, an estimated 0.85% of all Tweets are geo-tagged (Sloan and Morgan 2015). While this does not reflect to the share of users that regularly posts geotagged content (a percentage that could be higher), it does reveal that the trend of geotagging on Twitter is not very common. However, with an estimated 500 million tweets posted globally every day, this amounts to almost 4 million daily tweets from which to create a dataset (Sloan and Morgan 2015). Additionally, while many forms of social media may provide a larger dataset, in comparison to other platforms, Twitter's API stream is open access, allowing for easily replicable steps in the future.

Considering the above, this study aims to examine how Twitter data can be used for monitoring domestic and foreign tourism flows in case of Finland. We examine Twitter data to study visitation movements in Finland over the span of two years, October 2017 – October 2019. We evaluate the findings against official tourism statistics from Statistics Finland and finally evaluate the domestic tourism travelling distances. We use previously collected Twitter data by the Digital Geography Lab and our research is in line with GDPR – an act regulating the handling of personal data in the European Union. All personal information from the user profiles is excluded to protect against the re-identification of individuals.

Methods

The Twitter data was initially pre-processed by the *Digital Geography Lab* at the University of Helsinki (https://www.helsinki.fi/en/researchgroups/digital-geography-lab). The data was initially collected from the public Twitter API and then constructed by filtering out tweets without location information, leaving only geotagged tweets in the dataset. The study workflow is outline in Figure 1.

A heuristic programmatic approach by Massinen (2019) was utilized in provided data to find out the country of origin among Twitter users. The program was also used to determine the home municipalities of Finnish users, e.g. domestic visitors. However, due to uncertainties that arose when a user's tweets seemed to be coming evenly from two locations, the home country or home municipality remained undetermined. This was resolved by choosing the location randomly with a 50/50 probability for each location.

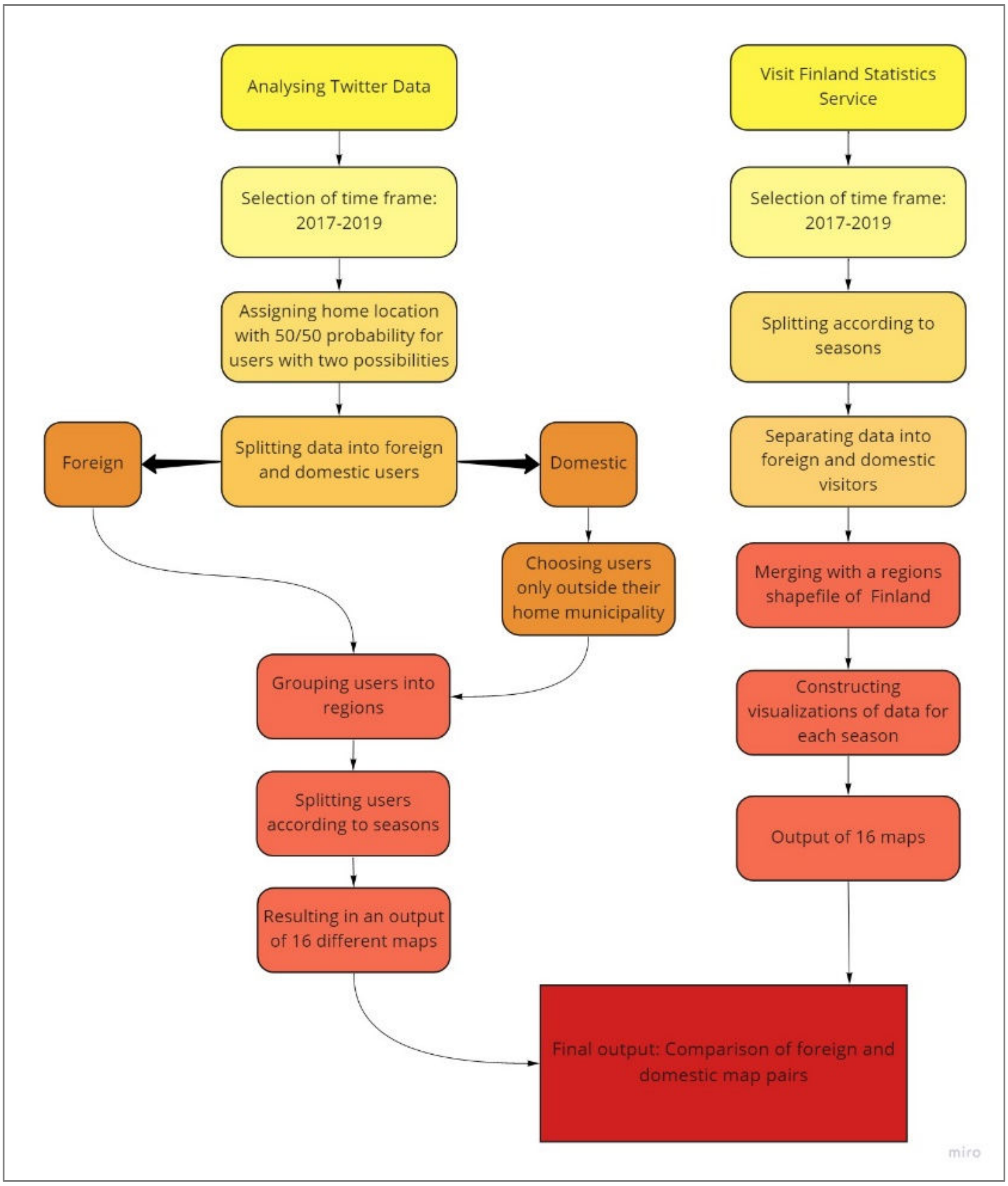


Figure 1. Workflow chart of methods

The users were separated into two data frames: foreign and domestic visitors. The data was narrowed to focus on the two most recent years accessible, 2017–2019. Domestic visitors were further filtered by only choosing those users tweeting outside their home municipality. Points for both were then grouped into the regions, split into seasons and compared to data from official statistics.

The official data is from Visit Finland's Statistics Service Rudolf with data from Statistics Finland and shows nights spent at accommodation per month by visitors' country of origin. It was available at a regional level and contained totals of domestic and foreign visitors. The data was first downloaded as an excel and split into seasons. The data was then saved as three separate csv files for domestic, foreign and all visitors and merged with a shapefile of the regions of Finland from The National Land Survey of Finland obtained from: https://avaa.tdata.fi/web/paituli/latauspalvelu.

Additional examinations were performed on domestic visotors traveling outside their home region. All twitter data from 2017–2019 was joined with the municipality information for each user. This was further modified with the addition of the region each municipality belonged to, and which users' tweet locations were outside of their home region. Distance travelled was then calculated by determining the distance between that user's home region's central point and the location of their tweet. Popularity of regions for domestic visitors according to their Twitter activity was determined by the number of unique user tweets in each region.

Results

Applicability of Twitter data for tourism

One of the research questions we wanted to investigate, was whether geotagged Twitter data can be used as a substitute or proxy for official tourist statistics. To investigate how similar the datasets are, we compare datasets regarding the overall picture of tourists in Finland. Table 1 show the ten largest visitor groups of both datasets by country of origin. The top ten visitor countries in both official and Twitter data are almost the same except for China and Estonia in the official data, and Spain and Italy in the Twitter data.

As Twitter has been blocked in China, it's absence in the twitter data is easily explained. The absence of Estonians from the Twitter data and somewhat low rank in the official data is surprising though it could be attributed to the type of stay. Due to their geographical proximity, they could be coming for shorter visits. The figure could be different if data from ports was used. Figures based on twitter data also convey the popularity of Twitter in the respective countries. Similarly to China, where Twitter is not used, there is a bigger share of tweeters from the US than there are hotel nights from the

USA. But in both datasets, Russians are the biggest group showing that the datasets do match quite well.

Given that the origin countries in both datasets are similar we want to check the validity of the geographical aspect of the data. That is, are the tweets corresponding to the locations that the official statistics show that are being visited. To investigate this, we divided the data into seasons, to allow for a wider time range. Figure 2 and 3 show how the Twitter data (left-side map) compares with the official statistics (right-side map) for the winter of 2017–2018 (December – February). Figure 2 shows Finnish visitors by region and which regions get the highest and lowest share of visitors, while the second figure shows the same but for foreign visitors.

Table 1. The comparison of foreign visitor distribution in Finland by origin country between the Official Data and Twitter data.

Country	% of visitors based on hotel nights	% of visitors based on twitter statistics
Russia	11,92%	8,25%
Germany	9,21%	2,34%
United Kingdom	8,36%	7,7%
Sweden	8,19%	5,6%
China	5,06%	0,32%
France	4,5%	1,54%
United States	4,27%	7,28%
Netherlands	3,63%	1,35%
Estonia	3,39%	0,8%
Japan	3,22%	2,5%
Norway	2,74%	0,9%
Spain	2,52%	3,18%
Italy	2,43%	1,77%
Turkey	0,42%	1,25%
Others	30%	25%

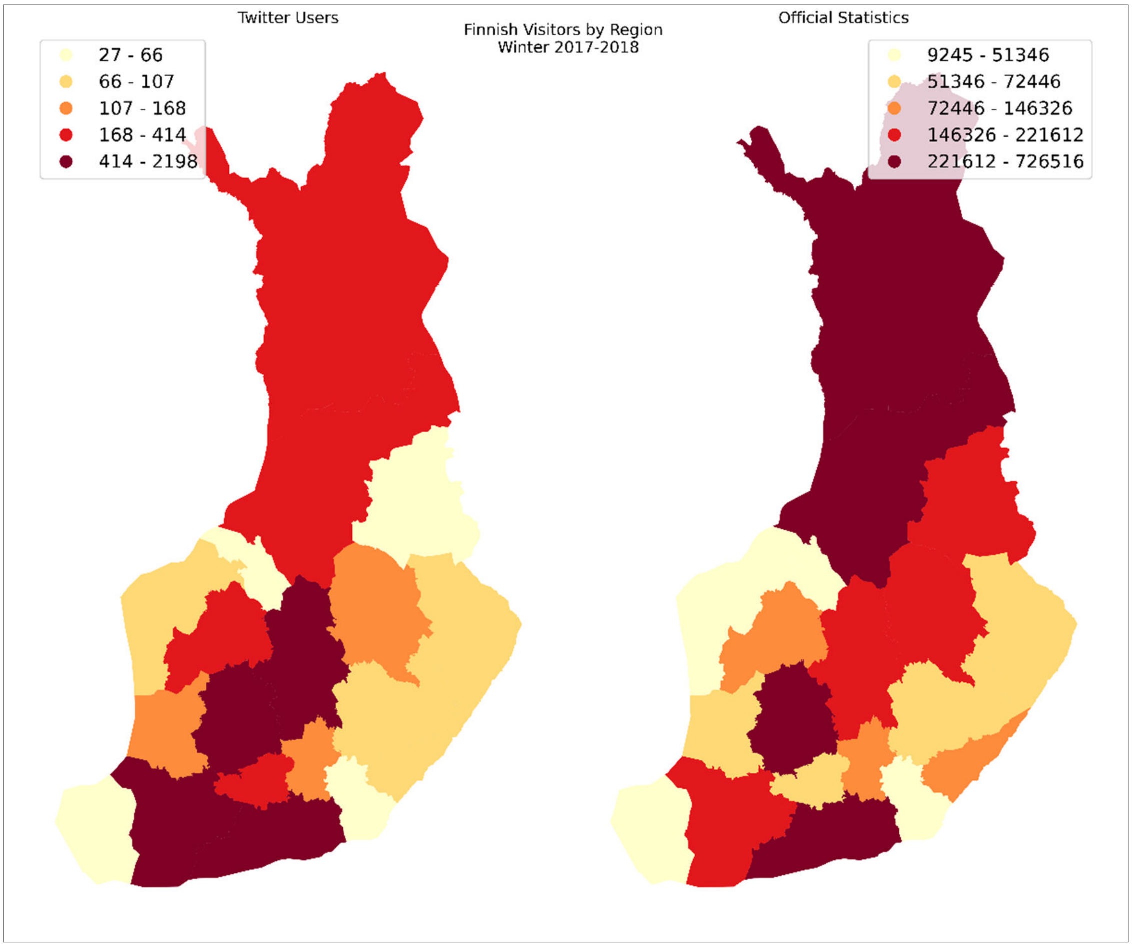


Figure 2. Share of Finnish visitors per region, Twitter data and Official Statistics for the winter 2017-2018. Correlation graph with a Pearsons correlation at 0.92.

The overall trend in both maps show that the Twitter data match up quite well to the official statistics. There are differences in the output maps, but one must remember the differences in input as well. Some of the reason for the differences may be attributed to the difference in collecting data.

The official statistics shows a night spent in the given region, while in the Twitter data one user might have entered one region, tweeted, and exited the region, all in the same day. Another difference might be attributed to how we in our analysis of the Twitter data have counted a visit. For the Finnish users we count a visit based on municipality level, so that one user can be visiting his or her own region. That will include also people tweeting in another municipality than where they live, e.g. if they commute to another municipality in the same region for work, and tweets while at work, that is counted as a visit. On the other hand, the same can happen in the official statistics, a person might stay at an accommodation in another municipality in the same region.

What these findings show is that it with relatively easy comparisons is possible to use the Twitter data at an overall level. There are still some hurdles that one need to overcome in the methodology to increase the validity and use case for this kind of social media data. How does one ensure the most representative sample in the data for example, and how can one to a larger degree ensure that the datasets represent the same activity?

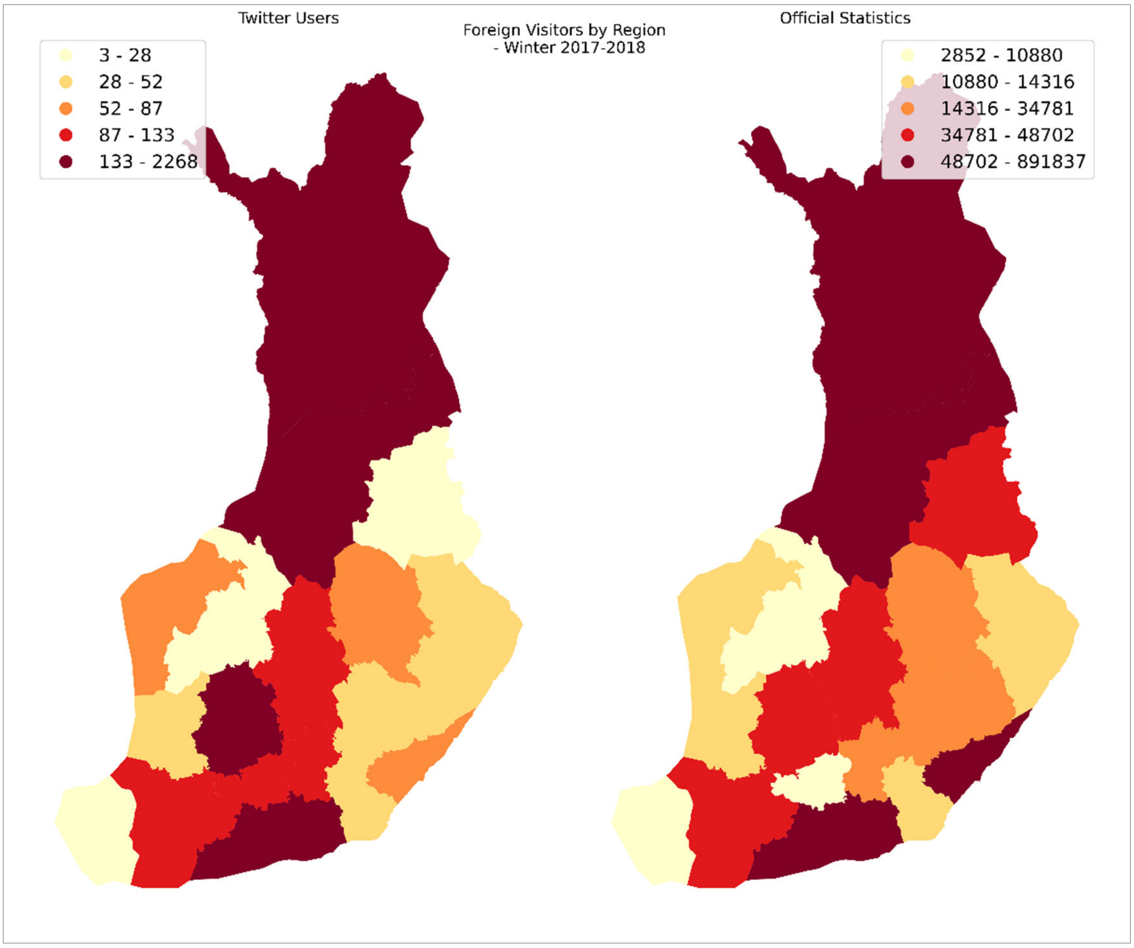
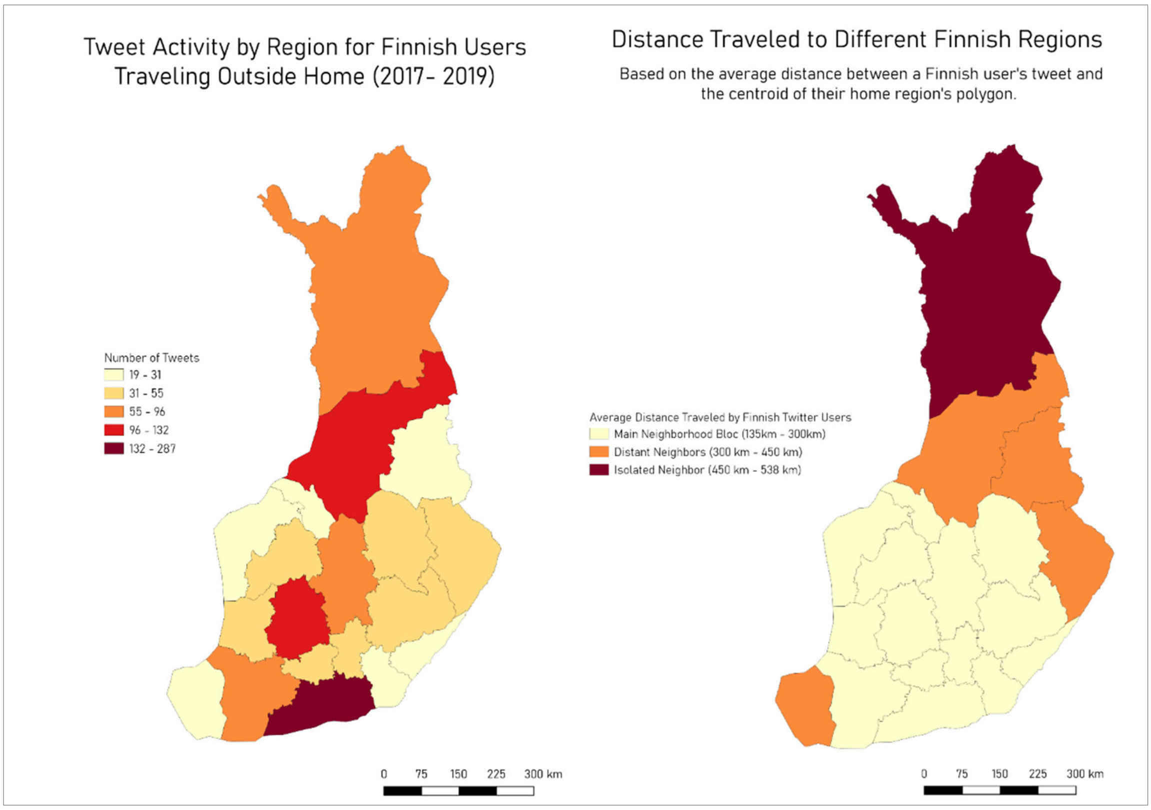


Figure 3. Share of foreign visitors per region, Twitter data and Official Statistics for the winter 2017-2018. Correlation graph with a Pearsons correlation at 0.81.



Figures 4. Average distance travelled by Finnish users outside their home region and Total number of tweets posted by Finnish users for 2017–2019.

Analysing domestic visitors travel distances

Of the active users found in Finland we were interested in further focusing on Finnish users to understand where Finnish people travel within their home country specifically through determining how far they travel and what regions are most popular in terms of tweet activity. The unique users tweeting outside their home region came out to be 1,237 users. By calculating each user's distance from home we are able to see what regions require the most amount of travel, and with a side-by-side comparison with region popularity according to tweet activity, we can see if the distance is worth the visit (Figure 4).

Though the numbers ranging over the two-year span for unique user tweets outside their home region are quite small, it can provide possible patterns for domestic tourist data especially when further divided into seasonal trends as previous figures demonstrate. The distance travelled reveals a natural result for regions requiring most amount of travel. Regions like Lapland that are isolated and farthest from the smaller regions found in Central Finland, naturally require a longer trip to visit.

Similarly, smaller regions in the centre require less travel due to their proximity to each other and size. Popularity of tweets by region do follow some natural trends especially with the popularity of Uusimaa, but new revelations of where Finnish people traveling outside their home region are most active online can also be concluded, like the region of Northern Ostrobothnia. While this region may require traveling over 300 km for the average Finnish user, it is one of the more popular regions according to the breaks set by a Natural Breaks classification for region popularity.

Discussion and conclusion

Overall, the results of this study prove the usability of Twitter data as a complimentary data for traditional data collection via registers and surveys in tourism research and statistics. Trends in the winter season demonstrate that the Lapland area receives a high share of both foreign and domestic visitors. Predictably, Uusimaa also receives a high share of visitors during this season. Distance travelled by domestic visitors to different regions across Finland resulted in a somewhat expected outcome, but further proves how social media can reveal further aspects about visitors' spatial behaviour. Certainly, we did not mitigate biases (e.g. by weighting) due to various demographic limitations of Twitter data. However, geographical trends in visitation patterns reflect strongly patterns from the official statistics recorded by the Visit Finland. Thus, this suggests that geotagged Twitter data can be used to monitor popular tourist destinations throughout Finland for each season of the year and to share this data with each respective region to give an idea of what each travel season looks like regarding domestic and foreign visitors.

This study used methodologically a simple approach and did not use different possibilities that advanced social media analytics could provide (Toivonen et al. 2019). While official tourism statistics are "official", one must remember that these are from incomplete accommodation statistics that are combined with modelled statistics based on a survey data, and these have its limitations. First, the official data excludes Airbnb or other alternative (non-commercial) accommodation. Second, Twitter data can reveal individual visitor movements within a destination whereas official data from Visit Finland cannot directly obtain. While nights in hotel can provide some reliable information for the number of tourists, it does not reveal, where and when they visit different activity locations. Social media data can reveal the popularity of certain tourist destinations and possible tourist transit routes.

Social media like Twitter data has it's biases and findings has to be assessed critically. Depending on an individual user's social media habits, it is never guaranteed how accurate the data for a certain user's trip is. A user might only post on a two-week long vacation, appearing statistically as though they only made a one-day visit. Also, some age groups or nationalities are more likely to use Twitter than others, thus a popular transit route based on tweets might only be true for one demographic. Nevertheless, representativeness issue is similar to survey research, and one solution can be classifying users to some categories and apply weighting technique.

Recommendations

In order to advance the more accurate assessment of tourism through social media, it would be recommended to follow official definitions of tourism as outlined by Eurostat (2014). Yet, people who fall under the group of frequent travellers who work in one country and live in another – cross-border workers – are not included in tourist statistics. Both places can be "*assimilated with the person's usual environment*", meaning their visitation into either country even with more nights spent in one over the other, does not qualify as a tourist (International Air Transport Association 2014). This is especially common from Estonia to Finland and Finland to Sweden. While we were interested in studying patterns on all visitation movements in Finland, our comparison to official tourism statistics could be improved by eliminating data figures for those visitors who do not officially qualify as tourists, per se.

Future studies should further involve more specific foreign visitor analysis based on a country of residence. More emphasis can be put on visitors travel movements and measuring travelled distances, and to differentiate specific visitor demographics groups. Finally, this study only focused on regions used in the official tourism statistics, however, Twitter data enables to conduct a municipality-level analysis or focus on some specific destinations such as nature reserves (see, Tenkanen et al. 2017).

References

Ahas, R., Aasa, A., Mark, Ü., Pae, T. & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, *28*(3), 898–910. https://doi.org/10.1016/j.tourman.2006.05.010

Chua, A., Servillo, L., Marcheggiani, E. & Moere, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management*, *57*, 295–310. https://doi.org/10.1016/j.tourman.2016.06.013

Eurostat (2014). *Methodological manual for tourism statistics: Version 1.3*. Publications Office.

Gheasi, M., Nijkamp, P. & Rietveld, P. (2011). Migration and tourist flows. In Á. Matias, P. Nijkamp & M. Sarmento (Eds.), *Tourism Economics* (pp. 111–126). Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2725-5_8

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V. & Minin, E. D. (2018). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters*, *11*(1), e12343. https://doi.org/10.1111/conl.12343

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3), 260–271.

Heikinheimo, V., Tenkanen, H., Bergroth, C., Järv, O., Hiippala, T., & Toivonen, T. (2020). Understanding the use of urban green spaces from user-generated geographic information. Landscape and Urban Planning, 201, 103845.

Kemp, S. (2020). Digital 2020: Global Digital Overview. Retrieved April 22, 2020, from https://datareportal.com/reports/digital-2020-global-digital-overview

International Air Transport Association (IATA) (2002). *General guidelines for using data on international air-passenger traffic for tourism analysis*. World Tourism Organization.

Massinen, S. (2019) *Modeling Cross-Border Mobility Using Geotagged Twitter in the Greater Region of Luxembourg*. MSc thesis retrieved April 22, 2020, from https://helda.helsinki.fi/handle/10138/306530

Miller, H., Goodchild, M. (2014) Data-Driven Geography *GeoJournal (2015)* 80:449–461.

M.Q, R., Célia, de, A., Cláudia Ribeiro, & Odete, F., Paula. (2019). *Handbook of Research on Social Media Applications for the Tourism and Hospitality Sector*. IGI Global.

Saluveer, E., Raun, J., Tiru, M., Altin, L., Kroon, J., Snitsarenko, T., Aasa, A., & Silm, S. (2020). Methodological framework for producing national tourism statistics from mobile positioning data. *Annals of Tourism Research*, *81*, 102895.

Sloan, L., & Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, *10*(11).

Statistics Service Rudolf. Accessed from: https://www.businessfinland.fi/suomalaisille-asiakkaille/palvelut/matkailun-edistaminen/tutkimukset-ja-tilastot/tilastopalvelu-rudolf/

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, *7*(1), 1–11.

The National Land Survey of Finland. Accessed from: https://avaa.tdata.fi/web/paituli/latauspalvelu

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. Biological Conservation, 233, 298–315.

Appendix 1: Users by country.

Country	**Users**
Finland	4856
Russia	255
United Kingdom	239
United States	224
Sweden	175
Spain	100
Japan	78
Germany	73
Italy	56
France	47
Netherlands	42
Turkey	37
Australia	29
Norway	28
Estonia	26
Canada	25
Brazil	23
Denmark	22
Switzerland	21
Belgium	19
Mexico	19
Thailand	19
India	16
Philippines	16
Indonesia	15
Singapore	14
Ireland	13
Latvia	12
United Arab Emirates	11
China	10
Poland	10
Czech Republic	9
Portugal	9
Belarus	8
Iceland	8
Malaysia	8
Ukraine	8
Chile	7
Austria	7
Greece	6
Hungary	6
South Africa	6
Colombia	5
Lithuania	5
Argentina, Armenia, Azerbaijan, Bahrain, Bulgaria, Costa Rica, Croatia, Cyprus, Dominican Republic, Egypt, El Salvador, Guatemala, Hong Kong, Iraq, Israel, Kenya, Kosovo, Kuwait, Malta, Morocco, New Zealand, Nigeria, Pakistan, Palestine, Paraguay, Qatar, Republic of Korea, Romania, Saudi Arabia, Serbia, Slovenia, Sri Lanka, Swaziland, Taiwan, Tanzania, Uruguay, Uzbekistan, Vietnam	< 5 each
Unknown	1643
Total	**8328**

Charlier, V., Neimry, E. & Muukkonen, P. (2020). Epidemics and Geographical Information System. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 17–25. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter II

Epidemics and Geographical Information System

Charlier, V., Neimry, V. & Muukkonen, P.

valentin.charlier@helsinki.fi, *University of Helsinki*
emile.neimry@helsinki.fi, *University of Helsinki*
petteri.muukkonen@helsinki.fi, *University of Helsinki*

Introduction

Progress about Geographical Information System (GIS) and methods have been significantly developed since the SARS-CoV epidemic of 2002/2003 and seasonal influenza. It has provided an improvement in the understanding of the dynamics and epidemiology as well as the way of responding to an epidemic. For centuries, the mapping has been considering by health professionals as a key role for the tracking of the epidemic (Kamel Boulos & Geraghty, 2020).

The importance of spatial analysis and the use of GIS in the field of health and the study of diseases has been reinforced by the emergence of COVID-19. This disease appeared in Wuhan (China) in December 2019 and then turned into a pandemic, forcing governments to establish measures to contain its spread, such as border closures and quarantine. This new disease has impacted the economic and the public health system by its quick spatial diffusion (Singhal, 2020).

Since there have been many advances and increases in data accessibility and software development, Geographical Information System (GIS) and spatial analysis have found new applications and uses, notably in the field of health and disease control (Kistermann et al., 2001; Boyda et al., 2019). Moreover, these advances in GIS technology allowed to study the spatial variation of disease and its association with the health care system and the environmental factors (Tanser & le Sueur, 2002; Nuvolone, 2011). The spatiotemporal component of diseases and the increasing interest of scientists in the use of GIS in public health shows the opportunities that GIS offers in the study and the management of diseases (Lyseen et al., 2014).

Use of GIS in public health and disease studies

GIS can be a powerful tool to understand and mitigate a disease by mapping the geographic distribution of disease and related it to the associated risk factors and the health services available. It can also provide a spatial analysis of the epidemic trends over space and time and the hotspot's location to organize health resources for prevention and treatment (Kistermann et al., 2001; Boyda et al., 2019). In fact, the mapping of spatial and temporal variations of diseases provided by the GIS allows authorities to plan and implement health measures where they are most needed and where the results will be most effective (Tanser & le Sueur, 2002).

One of the benefits of using GIS is the methodology it offers to deduce the spatial spread of a disease based on emission points. Indeed, environmental data affecting health (water, soil, air) are sometimes only available at specific points, so GIS interpolation techniques must be used to study the spread of diseases (Kistermann et al., 2001). In addition to establishing connectionsbetween different types of data (location, demographics, exposure, air quality, access to health care, etc.) the GIS allows analysis by buffering, geocoding, and mapping (Nuvolone, 2011).

Moreover, GIS can provide to health sector a lot of other benefits such as information and education of professionals and public people; reduce the cost of any sanitary actions using models and projections; strengthen decision-making from the local to the global level and continuously monitor and analyse changes in disease events. But it's not all, applications of GIS in the health sector can be introduced such as environmental health, surveillance of waterborne diseases, modelling exposure to risky areas (pollution, electromagnetic fields,…) and the analysis of the current disease policy and measures (ESRI, 2011; Shaw, 2012).

In the event of influenza or contagious disease, the health authorities may use the data collected at international airports for the purpose of assessing the health status of the passengers. Based on these data and the use of GIS (Geocoding) technologies the authorities can estimate the areas of exposure, assess the spread of the disease, and possibly contain its propagation (ESRI, 2011).

Moreover, GIS technology can support public interventions such as prioritize sites for vaccinations stations and community clinics, cancelling public events, closing public places (schools, office, recreation areas, …), or identifying relevant quarantine areas

(ESRI, 2011). To deal with infectious disease outbreaks, health authorities can use several applications of GIS technologies. For example, spatial analysis can be used to identify the source of the outbreak. The data provided by the GIS can be used as a resource for people to identify the closest care areas (hospitals, itinerary, time, ...). Moreover, the application of GIS technology has enabled Chinese authorities to select optimal sites for the construction of emergency treatment facilities during the onset of a COVID-19 outbreak (Kamel Boulos & Geraghty, 2020).

GIS technology can also be used as a preventive tool against diseases by assessing groundwater quality. Indeed, it allows a spatial analysis and mapping of groundwater components such as pH, ion concentration, and spatial distribution of pollutants. Furthermore, GIS can be used to solve water availability problems, prevent floods, and manage water resources from local to regional scales (Ketata et al., 2012).

Thus, there is a lot of use of GIS in the health sector and in disease management. For example, in Africa, GIS is used as a major tool to understand and manage contagious diseases such as malaria, tuberculosis, and the human immunodeficiency virus. GIS has been used to analyseand model the occurrence, seasonality, and transmission intensity of those diseases. Furthermore, the results obtained by this modelling can be combined with population data to assess population exposure and mortality risks. It can also be combined with climate data to estimate the impact of global warming on disease distribution, frequency, and intensity (Tanser & le Sueur, 2002).

Limitation of use of the GIS in the health domain

Despite the many advantages of using GIS in disease detection and prevention, there are many limitations and challenges for the future. First, there are some problems related to data concerns. Indeed, without adequate data, the accuracy of results in GIS cannot be relevant. In the domain of the diseases, there are specific problems areas such as how the disease data are reported and the mistakes data due to the movement of people (Sipe et al.,2003). The availability of data is also a current problem because there are many cases where digital data are not available or there is a lack of money to collect data. The availability of data faces other issues such as national security and confidentiality, especially in the sector of human health (Sipe et al., 2003; ESRI, 2011).

Second, there are limitations related to the GIS technology (GIS software) and due to the lack of knowledge and skills on the GIS of users. These limitations include a lack of qualified staff who does not have enough GIS training and skills that could lead to a misinterpretation of results (Sipe et al., 2003).

Third, GIS application such as geocoding can introduce errors and bias which could impact the results of a study. These issues can be created by several factors such as incomplete or inadequate data and human mistakes during the processing (Nuvolone, 2011).

Another problem is related to the dissemination of information on public health problems via social networks. Indeed, while the use of social networks can help promote public health strategies, it can also lead to the wide diffusion of information on personal data of people affected by a disease (Liang et al., 2019). Furthermore, the privacy and confidentiality restrictions of spatial data about health status and outcomes can create structural barriers to the adoption of GIS in public health measures (Shaw, 2012).

A relevant example of the limitation in using GIS is the case of malaria in Africa. Indeed, due to a lack of access to spatial data because of budget and infrastructure constraints, studies on some diseases lack relevant statistical analysis. The problem of available data is not only specific to the health sector but has all fields using GIS technology such as archaeology, ecology, or agroforestry. Improvements in GIS would help these regions by refining the accuracy of disease modelling techniques (Tanser & le Sueur, 2002).

The skills and training in GIS are also relevant in this example because most of the searchers in GIS applications in Africa are controlled by outsiders and not by African scientists who have knowledge of the socio-economic context. In order to be entirely effective, GIS must be introduced by searchers having both local knowledge on the area and technologi cal skills in spatial analysis (Tanser & le Sueur, 2002).

Example of GIS use: the case of COVID-19

Through interactive and near-real-time dashboards, GIS has been an important component of the information during the COVID-19 outbreak. For instance, there is the Johns Hopkins University's Center for Systems Science and Engineering (JHU CSSE) dashboard which has counted hundreds of millions of views, hence became the most viewed dashboard for the COVID-19 outbreak. Another example is the World Health Organization (WHO) dashboard which only takes confirmed cases by laboratories. The WHO dashboard also presents the progression of cases along time. A common aspect between the JHU CSSE and WHO dashboard is the importance of the optimization of the mobiles in order to maximize the potential number of informed people. On the other hand, there is HealthMap which analyses and maps data from online media sources such as Google New, social media, or validated alerts from the WHO. A specificity of HealthMap is the personal aspect of the information for the user thanks to his location. Indeed, it is possible to be informed about the nearby disease transmission risks at the user scale. As said before, the mobile represents a significant part of the information. The mobiles provide even more thanks to the locations and applications. For example, the geosocial app from China. This app exploits the data from the disease case records and the movement of people: if someone has been suspected or diagnosed to be infected, all the other users who have been close to him during the last two weeks (which corresponds to the incubation period of COVID-19) are informed. A system using almost the same functioning has been developed in Guangzhou Underground (China). Each passenger, when enters a metro carriage, must scan a QR code that is specific to the carriage. Thus, if someone is later diagnosed with coronavirus, the other passengers of the carriages will be informed (Kamel Boulos & Geraghty, 2020). These geosocial apps provide crucial information to the user, but the question of data privacy could be debated again.

The spreading of information is partially driven by social media. But when information is not correct, its spreading may continue. To slow down the spreading of misinformation, social media, and the WHO have collaborated (Kamel Boulos & Geraghty, 2020). Indeed, when a word about coronavirus is mentioned on social websites such as Facebook or YouTube, people have direct access to the WHO website.

The analysis of the worldwide transport pattern can be very useful to anticipate the high-risk places to be highly infected. By analysis of the connectivity between cities, the WorldPop tried to model the movement of people out of the epicentre (Wuhan) prior to its lockdown. Indeed, they first analysed the movement from Wuhan to other cities within China. Then, they analysed the potential cities in the world which are highly connected to these Chinese cities. Most of them are Asian with Bangkok in the first place. Melbourne and Los Angeles are the first cities out of Asia, at the 14th and 15th places respectively (Lai et al., 2020). Moreover, GIS can help to estimate the infection risks of COVID-19 of geographical areas depending on time (Al-Ahmadi et al., 2019) thanks to spatial statistics, e.g. Kulldorff's spatial scan statistics and associated cluster analyses.

Regarding to the GIS use and development, differences exist between countries. For the case of Pakistan, GIS tools have taken more importance due to COVID-19 outbreak even though there are strong limits about the data (accuracy of the location, facility, or data collection instrument). Their main use of GIS is to take appropriate actions in high -risk areas after detecting these (Sarwar et al., 2020). A study about India shows similar limits to the Pakistani case. However, thanks to interpolation, they managed to predict the COVID-19 spread pattern (Murugesan et al., 2020). About the United States, a web mapping platform has been developed to observe the results of the enforced social distancing thanks to mobility statistical patterns from smartphone location big data. Their goals are to raise awareness of people, have an impact on political decisions, and contribute to better community response to the pandemic. Two after the social distancing announcement, on average people of most of the states, followed the request of the government. Nevertheless, there are some limits to the methodology since social distancing does not directly imply reduced mobility (Gao et al., 2020). Mol lalo et al. (2020) model the COVID-19 incidence rate at the country level scale in the United States thanks to several variables that could explain its spatial variability such as environmental and socioeconomic variables. It turns out that the local models represent significantly more observations than this modelling. That kind of study could improve the anticipation of future outbreak development. Finally, GIS has also contributed to the analysis of the performance of the latest travel restrictions and border control measures (Wells et al., 2020).

Conclusion

Geographic Information Systems (GIS) is a relevant technical tool for spatial analysis which for several years has been increasingly used in many fields. Due to the spatial and temporal component of diseases, GIS has taken on great importance in the health field by allowing the prevention, understanding, and management of diseases and their spatial diffusion. GIS technology also makes it possible to analyse and monitor the quality of environmental factors affecting the health of inhabitants (soil, water, air). The use of GIS in the field of health also allows the management of current health structures and public health policies.

However, GIS technology can suffer from several flaws such as lack of data, lack of GIS training skills, geocoding errors. The privacy and confidentiality of personal data is also a structural restriction of the use of GIS technology.

The case of COVID-19 has shown several uses of GIS. The numbers of mapping dashboards through the internet and the interest of them have significantly increased. New geosocial apps provide accurate and crucial information but imply a debate about data privacy. The transport pattern around the world has been determining for the anticipation of contagion risk. Even though the advancement in GIS is different between countries, they all recognise its important uses.

Despite these shortcomings, GIS technology and spatial analysis are relevant tools for disease management, health policy decision-making and the prevention of future health challenges.

References

Al-Ahmadi, K., Alahmadi, S. & Al-Zahrani, A. (2019) Spatiotemporal clustering of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) incidence in Saudi Arabia, 2012–2019. *International Journal of Environmental Research and Public Health* 16(14), 2520. https://doi.org/10.3390/ijerph16142520

Boyda D.C., Holzman S.B., Berman A., Grabowski M.K. & Chang L.W. (2019) Geographic Information Systems, spatial analysis, and HIV in Africa: A scoping review. *PLoS ONE* 14(5), e0216388. https://doi.org/10.1371/journal.pone.0216388

Esri (2011) Geographic Information System and pandemic influenza planning and response. *An Esri White Paper*, February 2011. https://www.esri.com/library/whitepapers/pdfs/gis-and-pandemic-planning.pdf

Gao, S., Rao, J., Kang, Y., Liang, Y., & Kruse, J. (2020). Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special* 12(1), 16–26.

Kamel Boulos, M.N. & Geraghty, E.M. (2020) Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *International Journal of Health Geographics* 19(1), 8. https://doi.org/10.1186/s12942-020-00202-8

Ketata, M., Gueddari, M. & Bouhlila, R. (2012) Use of geographical information system and water quality index to assess groundwater quality in El Khairat deep aquifer (Enfidha, Central East Tunisia). *Arabian Journal of Geosciences* 5, 1379–1390. https://doi.org/10.1007/s12517-011-0292-9

Kistermann, T., Dangendorf, F. & Schweikart, J. (2001) New perspectives on the use of Geographical Information Systems (GIS) in environmental health sciences. *International Journal of Hygiene and Environmental Health* 205, 169–181. https://doi.org/10.1078/1438-4639-00145

Lai, S., Bogoch, I. I., Watts, A., Khan, K., Li, Z., & Tatem, A. (2020) Preliminary risk analysis of 2019 novel coronavirus spread within and beyond China. University of Southampton. https://www.worldpop.org/resources/docs/china/WorldPop-coronavirus-spread-risk-analysis-v1-25Jan.pdf

Liang, H., Fung, I.C., Tse, Z.T.H. et al. (2019) How did Ebola information spread on twitter: broadcasting or viral spreading? *BMC Public Health* 19, 438. https://doi.org/10.1186/s12889-019-6747-8

Lyseen, A. K., Nøhr, C., Sørensen, E. M., Gudes, O., Geraghty, E. M., … Shaw, N. T. (2014) A review and framework for categorizing current research and development in healthrelated Geographical Information Systems (GIS) studies. *Yearbook of Medical Informatics* 9(1), 110–124. https://dx.doi.org/10.15265%2FIY-2014-0008

Mollalo, A., Vahedi, B. & Rivera, K. M. (2020) GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of The Total Environment* 728, 138884. https://doi.org/10.1016/j.scitotenv.2020.138884

Murugesan, B., Karuppannan, S., Mengistie, A. T., Ranganathan, M., & Gopalakrishnan, G. (2020) Distribution and trend analysis of COVID-19 in India: geospatial approach. *Journal of Geographical Studies* 4(1), 1–9. https://doi.org/10.21523/gcj5.20040101

Nuvolone, D., Maggiore, R.d., Maio, S., Fresco, R., Baldacci, S., Carrozzi, L., Pistelli, F. & Viegi, G. (2011) Geographical information system and environmental epidemiology: a cross-sectional spatial analysis of the effects of traffic-related air pollution on population respiratory health. *Environmental Health* 10, 12. https://doi.org/10.1186/1476-069X-10-12

Sarwar, S., Waheed, R., Sarwar, S., & Khan, A. (2020) COVID-19 challenges to Pakistan: Is GIS analysis useful to draw solutions? *Science of The Total Environment* 730, 139089. https://doi.org/10.1016/j.scitotenv.2020.139089

Shaw, N.T. (2012) Geographical Information Systems and Health: Current State and Future Directions. *Healthcare Informatics Research* 18(2), 88–96. https://doi.org/10.4258/hir.2012.18.2.88

Singhal, T. (2020) A review of Coronavirus Disease-2019 (COVID-19). *The Indian Journal of Pediatrics* 87, 281–286. https://doi.org/10.1007/s12098-020-03263-6

Sipe, N.G. & Dale, P. (2003) Challenges in using geographic information systems (GIS) to understand and control malaria in Indonesia. *Malaria Journal* 2, 36. https://doi.org/10.1186/1475-2875-2-36

Tanser, F.C. & le Sueur, D. (2002) The application of geographical information systems to important public health problems in Africa. *International Journal of Health Geographics* 1, 4. https://doi.org/10.1186/1476-072X-1-4

Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., ... & Galvani, A. P. (2020) Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences* 117(13), 7504–7509. https://doi.org/10.1073/pnas.2002616117

Heittola, S., Koivisto, S., Ehnström, E. & Muukkonen, P. (2020). Combining Helsinki Region Travel Time Matrix with Lipas-database to analyse accessibility of sports facilities. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 26–38. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter III

Combining Helsinki Region Travel Time Matrix with Lipas-database to analyse accessibility of sports facilities

Heittola, S., Koivisto, S., Ehnström, E. & Muukkonen, P.*

suvi.heittola@helsinki.fi, University of Helsinki
sonja.koivisto@helsinki.fi, University of Helsinki
emil.ehnstrom@helsinki.fi, University of Helsinki
petteri.muukkonen@helsinki.fi, University of Helsinki

** Corresponding author petteri.muukkonen@helsinki.fi*

Abstract

This project aims to simplify the process of connecting sports facility data from Lipas database (University of Jyväskylä, 2020a) with the Helsinki Travel Time Matrix (Accessibility research group, n.d.). The Lipas database contains spatial data over various sports facilities in Finland and the Helsinki Travel Time Matrix is a collection of files organised as a grid with information about travel times in the Helsinki Metropolitan Area. With a connection between these two data sets, it will be possible to see the travel times to a certain sports facility category in the Helsinki region. Establishing a link between these datasets can be useful for individuals, researchers and planners. With this toolpack one can easily select a sports facility category and get a TIFF raster returned, with the travel times included. The development of a toolpack has been made with Python programming language and it is available on GitHub (https://github.com/petterimuukkonen/sportsfacilities).

Keywords: accessibility; sport facility; travel time; Helsinki Metropolitan area; geodata merging; automatization

Introduction

Accessibility to services and facilities is an important factor shaping the growth and spatial change in cities (Hasan et al. 2017). Spatial accessibility can be defined as the ability to reach goods, services and activities or as the degree to which a service or facility is accessible by as many people as possible (Litman 2010; Reggiani et al. 2015). The concept of accessibility has also physical, cultural, socio-economic and financial aspects but this study is focused on the geographical aspect (Karusisi et al. 2013).

Spatial accessibility to services can vary greatly within a metropolitan area and therefore the place of residence can either limit or enable our everyday activities (Higgs et al. 2015). In the case of sports facilities, the accessibility may affect the use rate of facilities and therefore have positive health impacts in the area (Karusisi et al. 2013). Furthermore, to prevent possible spatial marginalisation and health problem accumulation, it is important to ensure that all neighbourhoods have access to sports facilities and to investigate how the accessibility of facilities could be further improved with urban planning.

The accessibility research has developed greatly with the progress of Geographical Information Systems (GIS), tools and software (O'Sullivan et al. 2000). Many challenges still remain, since the accessibility varies according to the travel method, time of the day, area and rush hour patterns among other factors. Salonen and Toivonen (2013) have addressed this issue by creating comparable travel times for bike, public transport and car with a door-to-door approach that takes into account the rush hour, waiting and transfer times for public transport as well as time used for finding a parking place. The same approach is also used in this project to make the travel modes more comparable.

This project provides tools for accessibility research related to sports facilities in the Helsinki Metropolitan Area. Our aim is to automate a process of combining Finnish sport facility data (Lipas) with accessibility data of the Helsinki region area (Helsinki Region Travel Time Matrix). As a result of the process, a user gets a raster file of the accessibility of the chosen sport facilities in the Helsinki region and can visualise the results with static and interactive maps. This project is part of GIS project work course 2020 in University of Helsinki.

Data

Helsinki Travel Time Matrix and YKR grid

Helsinki Region Travel Time Matrix 2018 (later HRTTM) is a set of text-files that consists of calculated travel times and distances from the Helsinki metropolitan area. The data is built on SYKE (Finnish Environmental Institute) YKR grid that has 13 231 individual grid cells, with width and height of 250 m. The travel times have been alculated from each YKR grid cell centroid to each YKR grid cell centroid and saved into separate text-files. Each text-file represents travel times and distances from surrounding grid cells towards a single grid cell, so that the total count of text-files in this dataset is 13 231. For making the text files spatial data, the HRTTM data can be easily joined to a YKR grid file that has already been clipped to match the Helsinki Region area. The HRTTM dataset was calculated and collected in 2018 by MetropAccess-project and Accessibility Research Group in University of Helsinki and is the latest publication of Helsinki Region Travel Time Matrix datas ets at the time. It is licensed under a Creative Commons Attribution 4.0 International license, and can be used freely (Accessibility research group, n.d.).

Travel times are calculated separately for four travelling modes; public transport, private car, cycling and walking (Accessibility research group, n.d.; Tenkanen et al. 2020). These four travel modes are furthermore divided into 10 different travel methods using different travel speeds and times of a day.

Travel times for public transport and private car are calculated using a door-to-door approach. This approach takes into consideration the entire journey from starting point to the destination in different travel modes, such as walking into a bus stop or parking lot and also possible waiting times on the way. Travel times are calculated in two different times of a day: morning rush hour (08:00-09:00) and midday (12:00-13:00). In addition to this, travel times for private car are also calculated based on existing speed limits and in public transport there is an option of choosing door-to- door approach with or without the possible waiting time at home before leaving (Tenkanen et al. 2020).

Since personal characteristics of a cyclist influence the travel speed vastly the travel times for cycling is divided into fast and slow cycling based on Strava network travel speed data averages. Additional minute is added into cycling times referring to unlocking and locking the bike. A static walking speed of 70 metres per minute is used in all travel modes (Tenkanen et al., 2020).

Each txt-file consists of 18 attributes: 1) from_id, 2) to_id, 3) walk_t, 4) walk_d, 5) bike_f_t, 6) bike_s_t, 7) bike_d, 8) pt_r_tt, 9) pt_r_t, 10) pt_r_d, 11) pt_m_tt, 12) pt_m_t, 13) pt_m_d, 14) car_r_t, 15) car_r_d, 16) car_m_t, 17) car_m_d, 18) car_sl_t (Accessibility research group, n.d.). The first two attributes define from which YKRgrid cell to which YKR-grid cell the calculations have been taken from. The attributes contain a YKR-ID for locating the exact grid cell. The last 16 attributes (nro 3 to 18) contain the calculated travel times and distances in different travel modes. In the attribute names walk stands for walking, bike for cycling, car for private car and pt for public transportation. The latter letter specifications are explained in table 1. Nodata values are presented as -1 in the dataset (Accessibility research group, n.d.).

Table 1. Specification of the letters in attribute field names in Helsinki Region Travel Time Matrix 2018 data.

Letters in attribute field names	Definition
t	Travel time in minutes
tt	Travel time in minutes for public transport with waiting time at home before leaving
d	Travel distance in meters
f	Fast speed for cycling (19 km/h)
s	Slow speed for cycling (12 km/h)
r	Rush hour (08:00–09:00, 29.01.2018)
m	Midday (12.00–13:00, 29.01.2018)
sl	Travel speed according to speed limits

Lipas sport facility data

Lipas (https://www.lipas.fi/etusivu) is a national database of sport facilities in Finland. The database is managed by the Faculty of Sport and Health Sciences in the University of Jyväskylä and funded by Ministry of Education and Culture (University of Jyväskylä, 2020a). The data is uploaded to the database by municipality sport service workers or by private operators, such as utdoor unions and sport governing bodies (University of Jyväskylä, 2020b). Lipas data can be used through Liikuntapaikat.fi -application, by downloading ready-made datasets or by using an open source web map service (WMS), a web feature service (WFS) or a REST-interface (University of Jyväskylä, 2020c).

Lipas data is open source data that has been licenced by Creative Commons Attribution 4.0 International (CC BY 4.0) (University of Jyväskylä, 2020b).

The Lipas database contains information on sport facilities that includes maintained sport places, outdoor routes and parks (University of Jyväkylä, 2020a). A sport facility has to be publicly accessed, maintained regularly and equipped appropriately to be added into the database. A sport facility can also be for example a guiding point, maintenance building or an outdoor fireplace that is related to sports or outdoor activities. All in all, there were over 37 000 sport facilities in the dataset in spring 2019 (University of Jyväskylä, 2020b).

The sport facilities are grouped into sport facility types. There are eight (8) main types of sport facilities in the database (University of Jyväskylä, 2020b):

(1) Outdoor places and services (Virkistyskohteet ja palvelut in Finnish)
(2) Outdoor courts and sport parks (Ulkokentät ja liikuntapuistot)
(3) Indoor sport places (Sisäliikuntapaikat)
(4) Water sport places (Vesiliikuntapaikat)
(5) Terrain/outdoor sport places (Maastoliikuntapaikat)
(6) Boating, aviation and motor sports (Veneily, ilmailu ja moottoriurheilu)
(7) Animal sports (Eläinurheilu)
(8) Maintenance buildings (Huoltorakennukset)

These main types have been further on divided into more specific sport facility types. Many of the types have specific information that is only relevant in that sport facility. Therefore, these sport types differ from each other also based on attribute fields. However, there are few attribute fields that exist in all of these type groups. These are for example the sport facility type name in Finnish, Swedish and English, sport facility type code, the actual name of the facility and so on.

Methods and workflow for combining datasets

The Lipas sport facility data will be fetched from its open source Web Feature Service (WFS). This way it is possible to use Lipas data without storing it locally and always use the most updated data. The Lipas database includes features from all geometry types; points, lines and polygons. However, area and line features are somewhat problematic when it comes to accessibility analyses. One cannot be exactly sure from which point a person can truly access for example to a hiking route or a forest park area. Therefore, in this project we will be using only point based sport facility data.

The HRTTM 2018 and YKR grid data will be downloaded to a local repository and used from there through the process, since the data are stable releases and aren't updated regularly. The HRTTM and YKR grid data are openly available and can be downloaded from the data providers' website.

The actual process of combining Lipas data to HRTTM data consists of six separate steps (figure 1). First the Lipas sport facility data we will be obtained with a WFS request (step 1). After this Lipas data will be spatially joined to a YKR grid file to understand in which grid cells the sport facilities lie into (step 2). By this we will know the correct YKR grid cell ids that will help us to find out which HRTTM text-files are relevant in the outcoming maps.

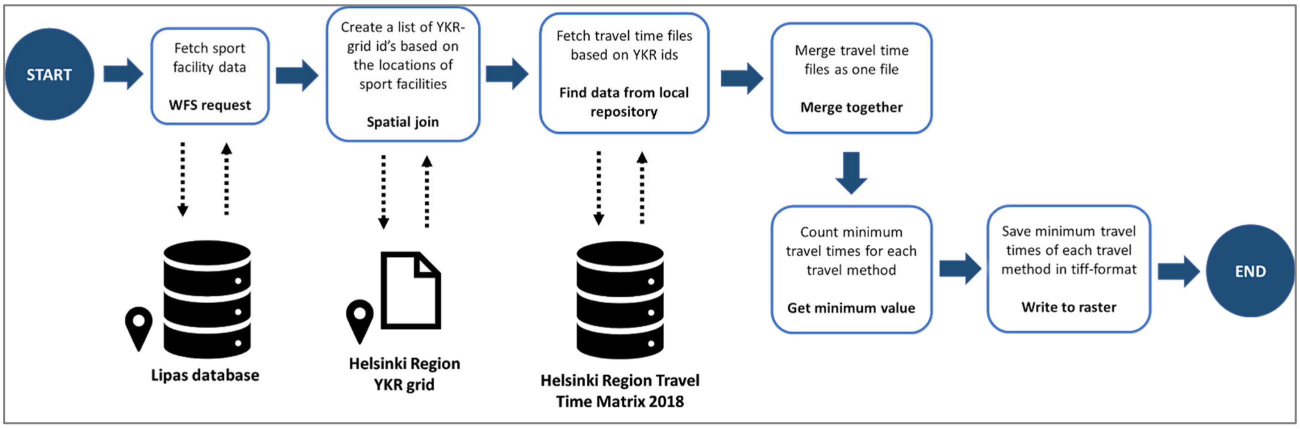


Figure 2. Simple workflow of the planned combining process.

When we have a list of YKR grid ids that correspond to the locations of the sport facilities we can obtain the correct HRTTM text-files from a local repository (step 3). If the chosen sport facility type has more than one sport facility there will most likely be more than one HRTTM file. These HRTTM files need to be merged as one file (step 4) and then the minimum travel times to any of the sport facilities can be calculated (step 5). After the minimum travel times have been calculated the data can be written into a raster file, visualized and used for example in research purposes (step 6).

We'll use Python programming language to build the automated process of combining the data and GitHub-platform as a version control service and shared platform for developing the code. GitHub will also be used in storing, developing and sharing the automated process further on.

Results

The final product of this project is a tool pack consisting of Python code functions that can be used for combining HRTTM data and Lipas data. The tool pack also includes functions that can be used for visualizing the results easily on a map. The tool pack includes eight functions (figure 2) in total that aim to simplify the process of combining HRTTM data with Lipas data. As the final output the tool pack can return the combination of the data as a TIFF raster file that includes travel times to chosen sports facilities with a particular travelling method. This raster file can then be used for further analysis or research and also be visualised in various GIS softwares.

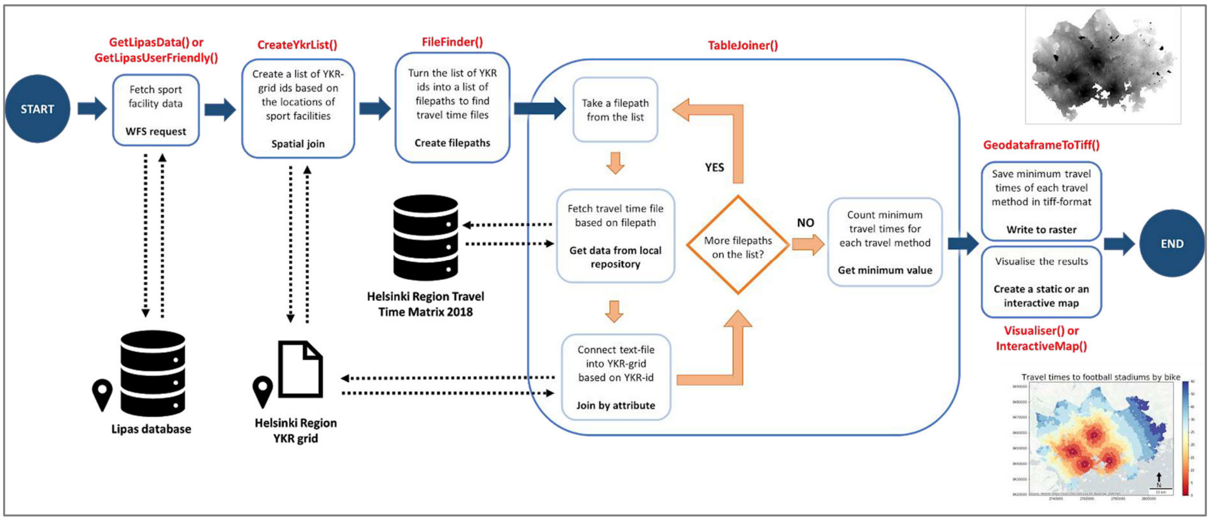


Figure 2. Workflow and the functions of the created tool pack.

The tool pack can be used by anyone, but since the tool pack is based on Python programming language, the assumption is, that it is most useful for developers, researchers and other people that have at least a brief understanding of programming. This tool pack is made with the geographical extent of the Helsinki Metropolitan Area and is therefore only useful for people interested in the particular area.

After downloading the HRTTM and YKR grid data from the data provider's website you can start using the functions found on Github. The sport facility data will be fetched from the Lipas database with a WFS request. This can be done with two functions: GetLipasData() or GetLipasUserFriendly(). By calling GetLipasData() – unction with two parameters, the sport facility type code and type name, the user will do a WFS request to the Lipas server, which will fetch the data. The type codes and type names can be found in a designated list in the GitHub repository or from Lipas database webpage (University of Jyväskylä, 2020d). As an example, one could choose to call GetLipasData("1350", "Jalkapallostadion") for fetching football stadiums. The GetLipasUserFriendly() -function is made for a more user-friendly approach. To use this function the user doesn't need to know in advance which type code or exact type name the sport facility has in Lipas database. As the function is called, it will list all available sport facility types for the user and the user can choose the one (s)he wants.

The fetched Lipas data needs to be spatially joined with the YKR grid, before it can be merged to the travel time matrix data. This can be done by using a function called CreateYkrList(). The function creates a list of YKR grid cell ids that have a sports facility inside them. As an input parameter user inserts the Lipas data that was fetched from the Lipas server. After the YKR id list is created a function called FileFinder() will turn the list of ids into a list of filepaths to the correct travel time files in the HRTTM dataset.

TableJoiner() function will fetch all the HRTTM text-files found based on the filepath list and turns the text-files into a geodataframe by joining the files into the YKR grid. After the files are joined as one, minimum travel times from each cell to the sport facility are calculated and a geodataframe is returned to the user.

After the table joining process, the data needs to be written into a raster file, so that it's easier to use. This is done by the function GeodataframeToTIFF() that takes a geodataframe (recently created with the TableJoiner()), sport facility type name and type code as input parameters. The function turns the input geodataframe of minimum travel times into raster format and returns a set of raster TIFF files for each of the ten different travel modes in HRTTM data. These raster files can then be easily used to visualize the accessibility of the sport facilities (figure 3).

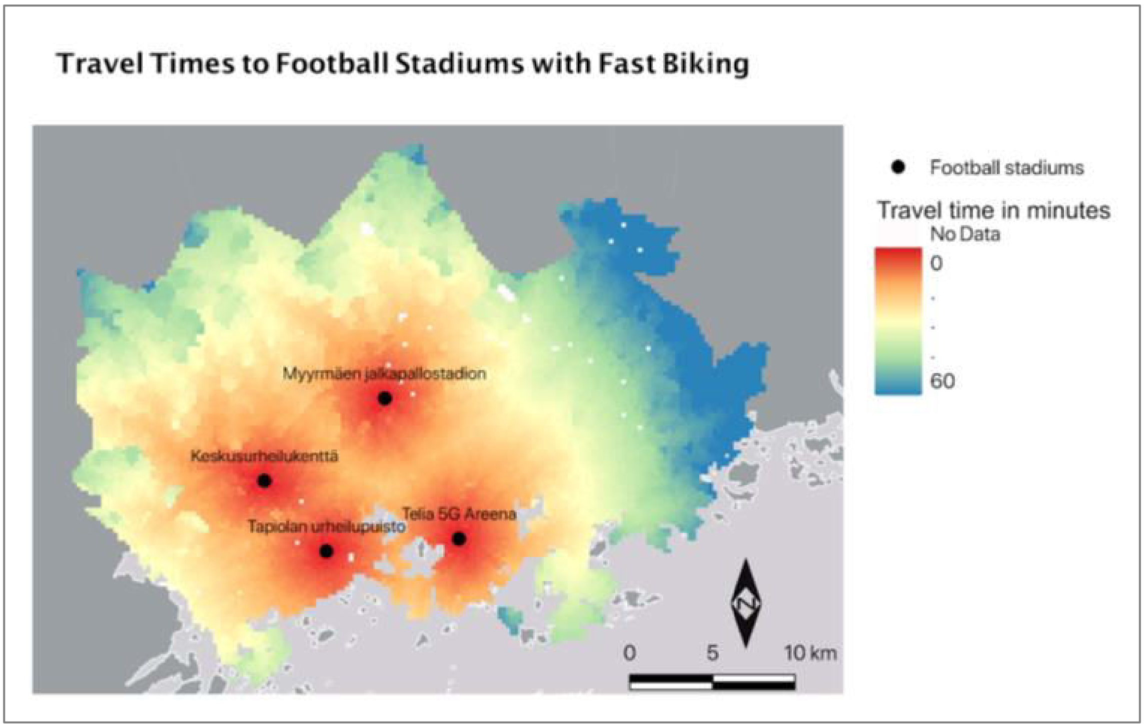


Figure 4. Simple visualisation of the raster output file in QGIS.

The Visualiser() function requires three parameters: a geodataframe with minimum travel times, the column name of the chosen travel method and the sport facility type name. The sport facility type name is only used for the output map title. Since this function has been designed to work only in English, it is recommended to insert the sport facility type name in English. As an output Visualiser () generates a map presentation as a png file and places it in the output folder. The map visualises the minimum travel times to the closest sport facility in chosen travel mode. The output file is a static image and is therefore easy to use in text documents and on paper (figure 4).

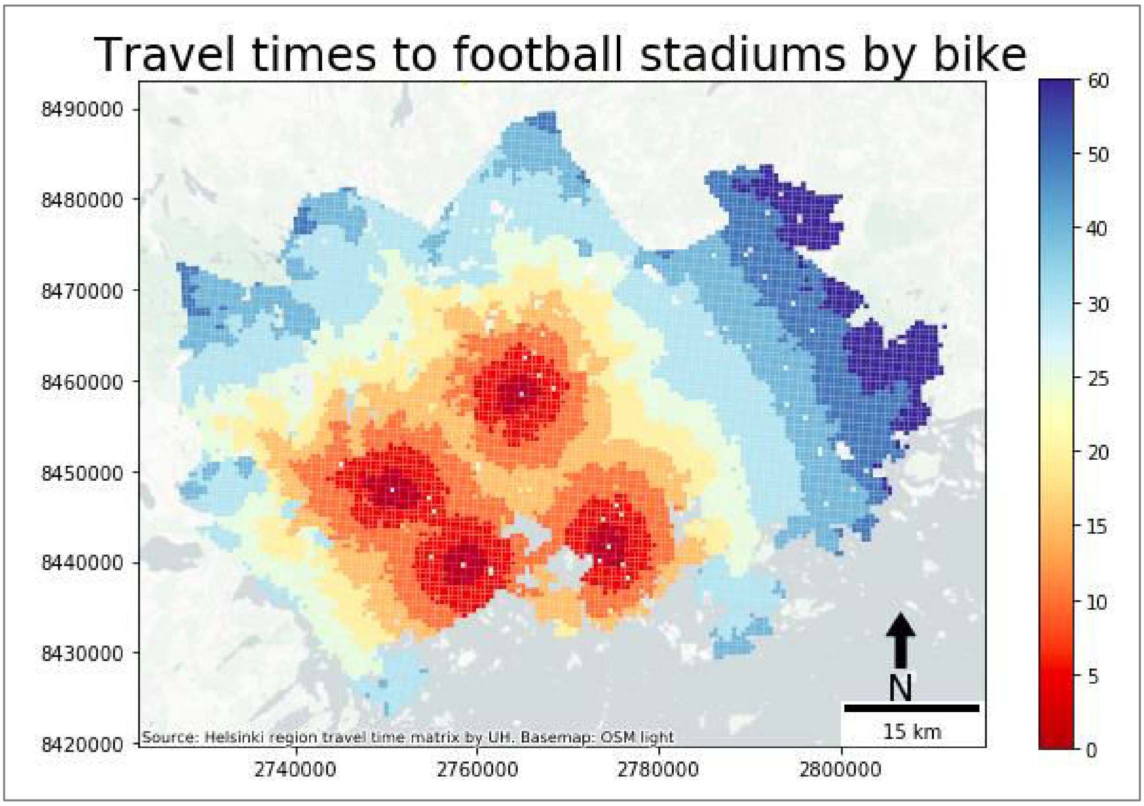


Figure 4. An example of what kind of output the Visualiser() function creates for travel times to football stadiums by bike.

A more suitable map for the web can be made with the function called InteractiveMap(). The function requires two parameters: one geodataframe and a column name (in other words the travel mode you wish to display). The InteractiveMap() returns an html file that contains an interactive choropleth map with the correct travel times for the chosen sports facility type. All the functions can be found on the GitHub page https://github.com/petterimuukkonen/sportsfacilities.

Discussion

The tool pack we present serves functional purposes, however, for someone who is not familiar with programming, learning how to use the tools can be overwhelming. Even for people who are familiar with Python programming probably need to refer to the documentation and demo material we have made in order to effectively use the functions. We see a need for a more user-friendly approach to our problem, where anyone could choose any set of sports facilities to be reached. This would make our work more accessible to the masses and our groundwork would become more useful. A solution to this would be producing a geodatabase containing the accessibility rasters for all travel methods for all sports facility types. This would be a kind of online library where all the files are stored and can be freely used for research purposes.

Currently, the geographical extent for using the tools we have created is quite limited. We are using the HRTTM as an input, and this type of accessibility data has not been created for different areas with the same format. Furthermore, the production of this type of accessibility data requires time resources and funding. Therefore, the approach is not feasible for extensive areas, at least with the current level of technology. If travel time matrices were created for other city regions in Finland, the tools could be easily extended to work for those areas as well. This project would not have been possible without up-to-date and quality data sources like HRTTM and Lipas database. If other countries provide and maintain similar databases, comparative research could be done and this can improve the understanding of spatial accessibility of sport facilities globally and the consequences.

We acknowledge that there are many possible points of improvement in our work. For a 2.0 version, if one would be made, we would add many functionalities which include at least the following:

In this version, the polygon and line features, like outdoor areas or orienteering terrains and jogging paths or ski tracks, are just filtered out. In the next version, the travel times could be calculated to the closest point of the polygon or polyline if it can be entered from any point. Another option would be to locate the enter points of routes or park areas and such, if that is even possible.

Currently, the user can only choose one sport facility subtype (like football fields) when fetching the data without modifying the function. For 2.0 version, the Lipas data fetching functions could be developed to fetch all subtype features inside a chosen sport facility maintype (like ballsport fields) or to choose multiple separate sport facility subtypes to be fetched at once, if the user wants so.

Currently, there is a slight bug in the HRTTM data, which results in NoData value in the YKR grid cell where the travel times have been calculated to. Therefore in some outcome raster files there are NoData values (-1) in the raster cells which represent the locations of sport facilities. These NoData values could be manually turned into 0 values, in order to distinguish it to be a destination point (where travel time is 0) and not actually missing data.

The Visualiser() function could be upgraded to have multiple languages and choose the language according to user input. A challenge here is the conjugation of words in Finnish.

Ideally, we would also host the travel time data (HRTTM and YKR grid) on an online platform where it is ready to be used with our demo notebooks and functions so that every user would not have to download and extract the files which can take some time.

Conclusions

Accessibility of sports facilities is an important topic that can affect the activity levels and even health of urban dwellers (Higgs et al. 2015; Karusisi et al. 2013). With this project, we have provided a toolset for further accessibility research by writing functions to combine the accessibility data and the sports facility data. With these tools, the user is able to choose facilities of their interest and assess the accessibility of those by creating maps and raster files for further processing and analysis in GIS softwares. Furthermore, different time considerations like rush hour and waiting time at home for public transport have been taken into account for the different travel methods and this makes them more realistic and comparable with each other (Accessibility research group, n.d.).

The functions we created can be run for different sport facilities inputs and the results can be gathered to an accessibility geodatabase that could be available online for other users. This was one of our goals, but we could not see it through due to restrictions during the pandemic. Freely accessible database could facilitate more accessibility research and make the results available for all, not just for those who handle the basics of Python.

References

Accessibility research group (n.d.). Helsinki Region Travel Time Matrix (2018). University of Helsinki. https://blogs.helsinki.fi/accessibility/helsinki-region-travel-timematrix-2018/ Read 4.3.2020.

Hasan, S., Wang, X., Khoo, Y.B. & Foliente, G. (2017). Accessibility and socio-economic development of human settlements. *PLoS ONE* 12(6). https://doi.org/10.1371/journal.pone.0179620

Higgs, G., Langford, M. & Norman, P. (2015). Accessibility to sport facilities in Wales: A GIS-based analysis of socio-economic variations in provision. *Geoforum* 62: 105–120. https://doi.org/10.1016/j.geoforum.2015.04.010

Karusisi, N., Thomas, F., Méline, J. & Chaix B. (2013). Spatial accessibility to specific sport facilities and corresponding sport practice: the RECORD Study. *International Journal of Behavioral Nutrition and Physical Activity* 10, 48. https://doi.org/10.1186/1479-5868-10-48

Litman, T. (2010). Accessibility. A Dictionary of Transport Analysis. 2010. pp. 1–3.

O'Sullivan, D., Morrison, A. & Shearer, J. (2000). Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. *International Journal of Geographical Information Science* 14(1), 85–104. https://doi.org/10.1080/136588100240976

Reggiani, A., Nijkamp, P. & Lanzi, D. (2015) Transport resilience and vulnerability: The role of connectivity. Transportation Research Part A: Policy and Practice 81, 4–15. https://doi.org/10.1016/j.tra.2014.12.012

Salonen, M. & Toivonen, T. (2013). Modelling travel time in urban networks: comparable measures for private car and public transport. *Journal of Transport Geography* 31, 143–153. https://doi.org/10.1016/j.jtrangeo.2013.06.011

Tenkanen, H. & Toivonen, T. (2020). Longitudinal spatial dataset on travel times and distances by different travel modes in Helsinki Region. Scientific Data 7, 77. https://doi.org/10.1038/s41597-020-0413-y

University of Jyväskylä (2020a). Lipas Liikuntapaikat.fi. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi. Read 17.3.2020.

University of Jyväskylä (2020b). Lipas-järjestelmän esittely. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi/esittely-2. Read 17.3.2020.

University of Jyväskylä (2020c). Rajapinnat ja ladattavat aineistot. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi/rajapinnat-jaladattavat-aineistot. Read 17.3.2020.

University of Jyväskylä (2020d). Lipas 2019 (Lipas 2.0) käyttöohjeet. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi/lipas-2019-2-0. Read 29.3.2020.

Laaksonen, I., Lammassaari, V., Torkko, J., Paarlahti, A. & Muukkonen, P. (2020). Geographical applications in virtual reality. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, 39‒45. Faculty of Science. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter IV

Geographical applications in virtual reality

Laaksonen, I., Lammassaari, V., Torkko, J., Paarlahti, A. & Muukkonen, P.

iivari.laaksonen@helsinki.fi, University of Helsinki
valtteri.lammassaari@helsinki.fi, University of Helsinki
jussi.torkko@helsinki.fi, University of Helsinki
arttu.paarlahti@helsinki.fi, University of Helsinki
petteri.muukkonen@helsinki.fi, University of Helsinki

Abstract

The aim of study was to create a brief literature review about virtual reality and its geographical applications. This article contains introductions to LiDAR 3D-data with a focus on airborne and terrestrial laser scanning, and a historical overview on virtual reality and its geographical applications. This chapter is a literature review about the use of virtual reality (VR) and its geographical applications. The concept of VR has become very topical to geography students and teachers at the University of Helsinki due to the newly introduced virtual reality cabins and uses of those in the department's research projects. This article provides an introduction on the use of virtual reality in education.

Keywords: 3D data, Virtual Reality, VR, Airborne laser scanning, ALS, Terrestrial laser scanning, TLS, LiDAR, Education, teaching

3D-data from airborne laser scanning and terrestrial laser scanning

In geographical applications, the most prominent method to collect 3D-data has been LiDAR. Two the most common applications for LiDAR technology are airborne and terrestrial laser scanning. In this section, we introduce two methods for collecting 3D-data in remote sensing: airborne (ALS) and terrestrial laser scanning (TLS). These approaches utilize LiDAR (Light Detection and Ranging) technology, which uses light in the form of a laser pulse to measure distance to objects. LiDAR methods have the potential to collect millions of measurable survey points in minutes. The final product for ALS and TLS is typically a georeferenced point cloud that have many applications is virtual reality, for instance.

With the LiDAR, the distance from the sensor to the target is measured based on the time between the emission of the pulse and the detection of the echo that is backscattered to the sensor (Hyyppä et al. 2008). The sensor emits multiple laser pulses, the returns of which are classified. Multiple returns provide useful information that can help assess forest structure, for example. The first returns can be assumed to echo from the top of the canopy, while the last returns most likely originate from the ground level.

TLS is a LiDAR approach where the scanner is attached to a mobile platform, which can be transported wherever the target is located. TLS has the potential to provide very dense point clouds and the applications are practically endless. One can, for example, apply the method to archaeology, urban modelling, and forestry. However, the approach is expensive and requires a lot of funding. Still, the TLS method is very interesting, especially, from the point of view of virtual reality applications. The following sections go deeper into the concept of virtual reality and its applications in scientific research and teaching.

Virtual reality (VR)

History

Virtual reality (VR) groups together several technical inventions, which try to allow a user to experience simulated environments. Yung and Khoo-Lattimore (2017) conclude that virtual reality uses navigable and interactable 3D environments to achieve real-time simulations for user's different senses. They also found out that there are three defining elements to VR. First being able to look around in the simulated environment, usually by using a head mounted display (HMD). Second one is immersion, though suspension of belief and by having objects physically represented as in sensible objects in the virtual world. The third one is interactivity, achieved by having some sort of control over the experience.

According to Cipresso et al. (2018), a virtual reality system can be categorized by the degree of immersion. There are non-immersive systems, which consist mainly of a desktop and are simple and the most affordable. Secondly, semi-immersive systems such as stereo images of a 3D scene. The immersive systems are what most people see virtual reality as nowadays. They usually consist of an HMD, along with audio and haptic devices.

The history of VR began in the 1960's (figure 2) with Sensorama, a multimodal experience of 3D scenes and smells that could be considered the first VR device (Omer et al. 2019). The first system to include an HMD and head tracking was the Sword of Damocles. Head tracking in the system was explored with both mechanical arms and ultrasonic waves (Sutherland 1968).

Cipresso et al. (2018) has mentioned that during the commercialization of VR that started in the 1980's, appeared the first instances of what we could now consider as hand tracking or haptic devices. Data Glove, a sensory device was launched in 1985 and could measure the flexion of fingers along with their orientation and position. Binocular-Omni-Orientational Monitor (Boom) launched in the late 1980's allowed moving and broad virtual environments, along with mechanical hand tracking.

Since 2000's, improvements in processing speeds and the general hardware sizes have allowed companies to produce more consumer orientated VR devices (Omer et al. 2019). Especially video game companies have advanced VR to have higher field of views along with lower latencies. The development has also allowed the introduction of new sensors, for example eye-tracking systems (Cipresso et al. 2018). Also, the inclusion of smartphones into VR systems has allowed a wider userbase for the devices. Modern HMD systems can be divided into three categories: the first and the second being smartphone using headsets with or without additional functionalities. The third category includes headsets with their own graphics, specifically designed to be used without a smartphone (McMillan et al. 2017). They are the ones that provide a total immersion in a virtual environment when used along with other sensory equipment.

LiDAR and virtual reality

When LiDAR and especially TLS produced data is used within a virtual environment, new approaches can be found. When using an HMD, the dataset can be projected to the user based on his/her head orientation and position in real time. VR allows the visualization, integration, manipulation and querying of geospatial data (Zhao et al. 2019). Kreylos et al. (2006) found supporting evidence to their hypothesis that visualizing data in VR allows for more accurate and confident observations in less time. The immersion provided by an HMD is essential for identifying quality problems in the LiDAR data and allows workflow to be more efficient and accurate (Kreylos et al.

2008). In another study, the participants thought that using VR to viewpoint clouds allowed them to have a better idea of depth and the data's real-life implications. They also thought it to be especially beneficial to be immersed and to be able to manipulate the data in real time (Burwell, Jarvis & Tansey 2012).

Using VR to viewpoint clouds isn't totally problem free. In the Burwell et al. (2012) study, some participants also seemed to experience an information overload. This problem was due to the data having a fixed level of detail in all scales. Also viewing data in VR requires high end hardware: to have an immersive and intuitive visualization the system needs to be able to support 48–60 stereoscopic frames per second (Kreyolos et al. 2008). Kreylos et al. (2006) also believe that having the ability to manipulate the data directly in the visualization is crucial for the usage, which increases the cost of a system due to the need for hand tracking or other censors.

Zhao et al. (2019) have theorized a workflow for exporting LiDAR data into a VR system. Currently a game engine is needed to run a simulation in VR. One of the most common ones is Unity (Unity 2020), which is a cross-platform game engine (Navarro et al. 2018), but others with similar functionalities are available as well. Unity has the advantage of a free license and an accompanying development environment. It has been used together with LiDAR data on multiple studies (e. g. Mures et al. 2016, Garry et al. 2018, Navarro et al. 2018).

The VR engine used in Unity does not allow point clouds to be imported straight away, but instead requires data in a mesh format, which is a combination of vertices, edges and faces (Navarro et al. 2018). Before the mesh creation, the points can have shading values or RGB values assigned to them (Zhao et al. 2019). To create a mesh from the point cloud, the data first needs to be generalized, to reduce the amount of points. After the reduction, it is triangulated and imported to unity as a mesh. Finally, it is used with a Unity based program or script (Navarro et al. 2018).

Virtual reality in education

Virtual reality has been used in teaching since 1960's (Kavanagh et al. 2017). In the beginning they we mostly used in different kinds of pilot simulators, but slowly they have spread to other areas of education and training too. But even though it has become more common especially in higher education, its role is still limited in specific questions and subjects rather than being for used basic teaching method. Naturally studies what are originally based on the platforms of virtual reality are done in it, but also when tasks and practices which are too expensive, dangerous or difficult to perform in traditional ways. So, the reasons to for VR to spread is the possibilities it serves, safe environment and its cost-effectiveness on some occasion.

However, there are also problems in VR. According to Kavanagh et al. (2017) software usability is the most common issue, covering 48.6% of reported problems in VR use according to their studies. New software requires its own studying before VR itself can be used for studying other subjects. So at least the basics mechanics of the software need to be managed.

We see that in geography VR has several different possibilities to be used. In urban and regional planning, 3D pictures are used to demonstrate and observe plans. For example, in complementary building it's fast way to examine heights and shadows of building in right scale, and with 3D observing the picture is much more realistic than in 2D. Different kind of 3D softwares are also fast and flexible way to compare different solutions in placing buildings and structures without need to make the whole plan from the beginning.

Again, in physical geography getting to know your study area beforehand is easier and cost-effective compared to travelling to location. That is especially in distant locations and difficult terrain or wide areas. The benefits of VR can also be exploited in examining geomorphological formations and studying them. The process of erosion and the formation of different kinds topography are also easy present via VR.

Conclusion

This review shows that virtual reality (VR) can have prolific applications in the science of geography. LiDAR data, especially, has various applications to be used in the Department of Geosciences and Geography. The applications for virtual reality are practically endless and this report shows that especially educational applications are viable.

References

Burwell, C., Jarvis, C. & Tansey, K. (2012). The potential for using 3D visualization for data exploration, error correction and analysis of LiDAR point clouds. *Remote Sensing Letters, 3*(6), 481–490.

Cipresso, P., Giglioli, I. A. C., Raya, M. A., & Riva, G. (2018). The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology, 9*, 2086.

Garry, W. B., Ames, T. J., Bradnt, M. A., Slocum, S., Grubb, T. G. & Heldmann, J. L. (2018). Virtual analog environments: Exploring LiDAR data. *49th Lunar and Planetary Science Conference 2018*.

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X. & Maltamo, M. (2008). Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. *International Journal of Remote Sensing*, *29*(5), 1339–1366.

Kavanagh, S., Luxton-Reilly, A., Wuensche, B. & Plimmer, B. (2017). A systematic review of virtual reality in education. *Themes in Science and Technology Education*, 10(2), 85–119.

Kreylos, O. & Bawden, G. (2008). Immersive visualization and analysis of LiDAR data. *Advances in Visual Computing,* 846–855.

Kreylos, O., Bawden, G., Bernardin, T., Billen, M., Cowgill, E., Gold, R., . . . Sumner, D. (2006). Enabling scientific workflows in virtual reality. *Proceedings - VRCIA 2006: ACM International Conference on Virtual Reality Continuum and its Applications*

McMillan, K., Flood, K. & Glaeser, R. (2017). Virtual reality, augmented reality, mixed reality, and the marine conservation movement. *Aquatic Conservation: Marine and Freshwater Ecosystems, 27*, 162–168.

Mures, O., Jaspe, A., Padrón, E. & Rabuñal, J. (2016). Virtual reality and point-based rendering in architecture and heritage. *Handbook of Research on Visual Computing and Emerging Geometrical Design Tools,* 549–565.

Navarro, F., Fdez, J., Garzon, M., Roldán Gómez, J. & Barrientos, A. (2018). Integrating 3D reconstruction and virtual reality: A new approach for immersive teleoperation. *ROBOT 2017: Third Iberian Robotics Conference*, 606–616.

Omer, M., Margetts, L., Hadi Mosleh, M., Hewitt, S. & Parwaiz, M. (2019). Use of gaming technology to bring bridge inspection to the office. *Structure and Infrastructure Engineering, 15*(10), 1292–1307.

Omieno, K. K., Wabwoba, F. & Matoke, N. (2013). Virtual reality in education: trends and issues. *International Journal of Computers & Technology*, 4(1), 38-43.

Sutherland, I. E. (1968). A head-mounted three dimensional display. *AFIPS Conference Proceedings (1968) 33, I.*

Unity. (2020). Unity for all. Retrieved from https://unity.com/. Accessed 9.5.2020

White, J.C., Wulder, M.A.; Varhola, A.; Vastaranta, M.; Coops, N.C., Cook, B.D., Pitt, D. & Woods, M. (2013). A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. Natural Resources Canada, Canadian Forest Service, Canadian Wood Fibre Centre, Victoria, BC. Information Report FI-X-010.

Yung, R. & Khoo-Lattimore, C. (2019). New realities: A systematic literature review on virtual reality and augmented reality in tourism research. *Current Issues in Tourism, 22*(17), 2056–2081.

Zhao, J., Wallgrün, J. O., LaFemina, P. C., Normandeau, J. & Klippel, A. (2019). Harnessing the power of immersive virtual reality - visualization and analysis of 3D earth science data sets. *Geo-Spatial Information Science, 22*(4), 237–250.

Ruohio, P., Stevenson, R., Muukkonen, P. & Aalto, J. (2020). Compiling a tundra plant species data set. In Muukkonen, P. (Ed.): Examples and progress in geodata science, pp. 46–52. Department of Geosciences and Geography C19. Helsinki: University of Helsinki.

Chapter V

Compiling a tundra plant species data set

Ruohio, P., Stevenson, R., Muukkonen, P. & Aalto, J.*

petri.ruohio@helsinki.fi, *University of Helsinki*
rohan.stevenson@helsinki.fi, *University of Helsinki*
petteri.muukkonen@helsinki.fi, *University of Helsinki*
juha.aalto@fmi.fi, *Finnish Meteorological Institute*

* *Corresponding author* juha.aalto@fmi.fi

Abstract

Our task was to modify and compile a tundra plant-species dataset related to the research project *Microclimatic Heterogeneity Diverts Arctic Change Trajectories*. In total, there were 120 separate data sheets for 120 sample sites in the Kilpisjärvi area. In addition, there were 80 older sample sites. These plant observation files were compiled together to allow for further analyses using the data. This article contains a metadata description of the final output, description of the project workflow and some example maps and graphs.

Keywords: Arctic, Microclimate, Tundra plant-species

Introduction

This article presents how a dataset containing information about tundra plant-species in the Kilpisjärvi region was compiled to improve the usability of the collected data. The core of the article is a metadata description of the new dataset. Additionally, the workflow of the process and the possible uses of the dataset are briefly addressed.

The data has been collected as a part of a project *Microclimatic Heterogeneity Diverts Arctic Change Trajectories* - a research project funded by the University of Helsinki. The project is the first to show that microclimatic heterogeneity is the key in understanding Arctic change. The project addresses three specific objects: 1) developing new methodologies to realistically model the key microclimatic parameters (near-surface temperature, wind speed, soil moisture, snow) at high spatial and temporal resolution, 2) quantifying the magnitude of microclimatic decoupling in a topographically-rich Arctic

Ruohio, P., Stevenson, R., Muukkonen, P. & Aalto, J. (2020). Compiling a tundra plant species data set. In Muukkonen, P. (Ed.): Examples and progress in geodata science, pp. 46–52. Department of Geosciences and Geography C19. Helsinki: University of Helsinki.

landscape and examining its importance in buffering changing macroclimatic conditions, and 3) re-evaluating the consequences of climate change on multiple Arctic systems (vegetation dynamics, frost-induced geomorphology, permafrost) and surface-atmosphere feedbacks (carbon cycling and albedo). The dataset compiled during the project course will be combined with other data sources to achieve the aforementioned targets.

In total, there were 120 separate data sheets for 120 sample sites in the Kilpisjärvi area. These sites are a part of a larger study design of 200 sites, of which 80 sites were have been sampled earlier. The exact locations of the project sites are presented in Figures 1 and 2.

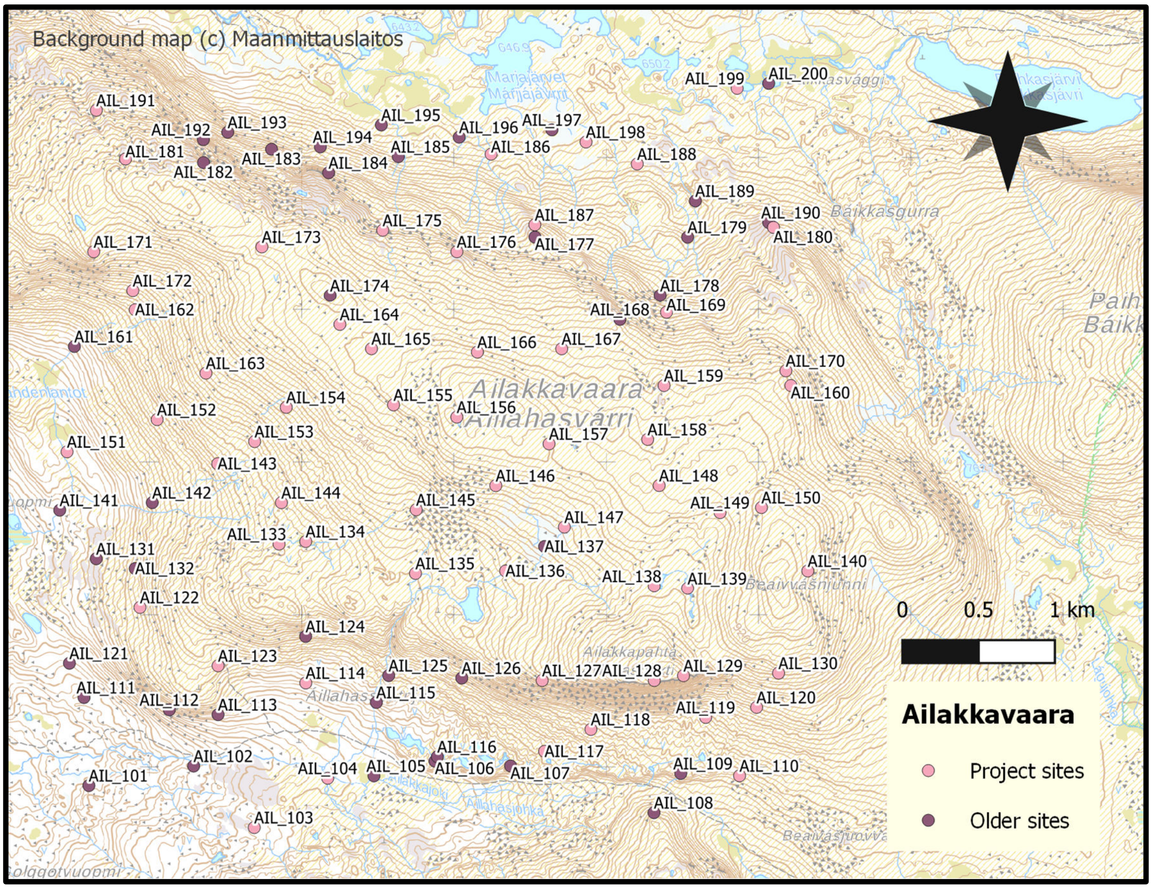


Figure 1. Study sites at Ailakkavaara, Kilpisjärvi.

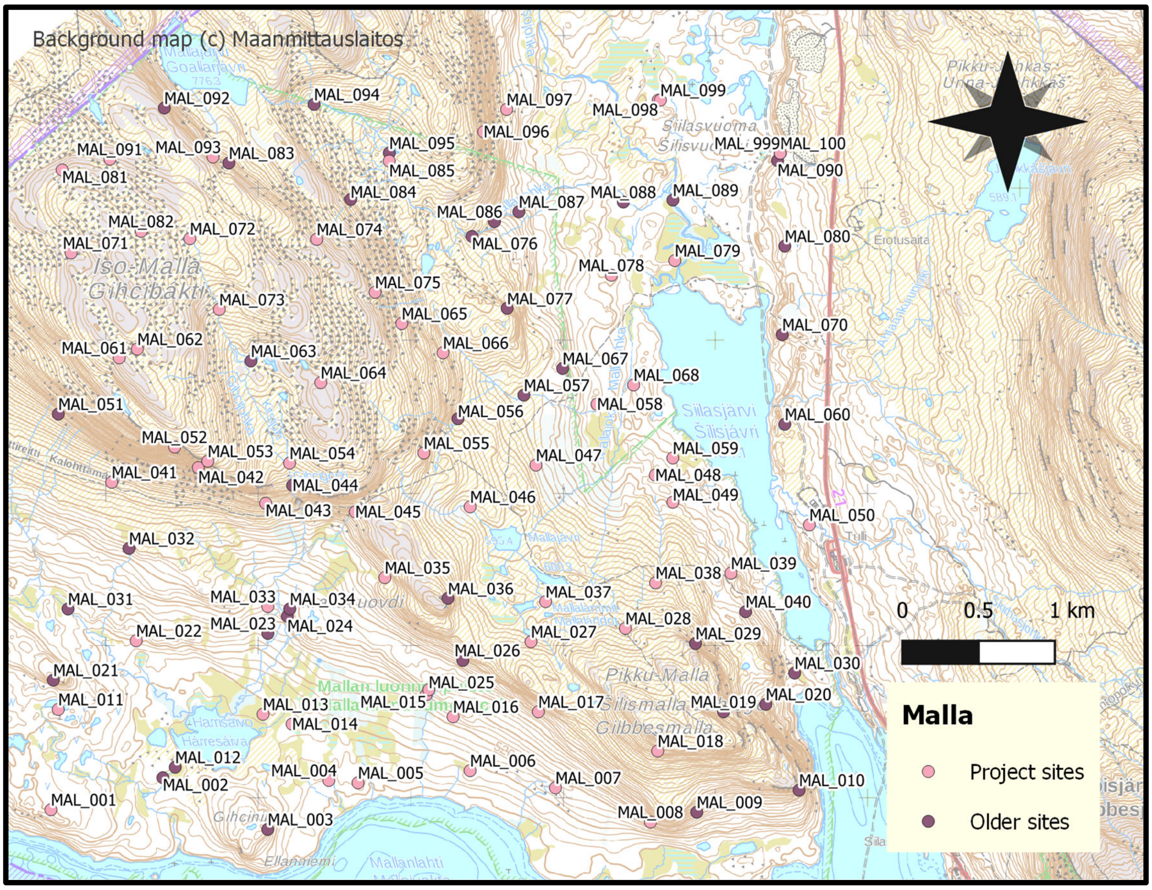


Figure 2. Study sites at Mallatunturit, Kilpisjärvi.

Metadata description

The data comprises three tables. Table 1 is a long-format table indicating the percentage cover of each species present in a 1x1 metre area as well as the phenological state of each species present. It comprises the following four variables:

- name = character, site name (massif + id)
- species_name = character
- cover = numeric (percentage)
- pheno_state = character (flowering, vegetative or n/a if no species present)

Table 2 is another long-format table indicating the presence or absence of the species from Table 1 in a 2x2 metre area. It comprises the following three variables:

- name = character, site name (massif + id)
- species_name = character
- cover = numeric (present=1/absent=0)

Table 3 is a wide-format table describing the general environmental characteristics of each site. It comprises the following variables:

- site_id = character (massif + id)
- habitat = character, habitat type of site
- avr_height_vegetation = numeric, average height of all vegetation (centimetres)
- max_height_vegetation = numeric, maximum height of all vegetation (centimetres)
- haylike_average_height = numeric, average height of haylike vegetation (centimetres)
- other_avr_height = numeric, average height of other vegetation (centimetres)
- haylike_max_height = numeric, maximum height of haylike vegetation (centimetres)
- others_max_height = numeric, maximum height of other vegetation (centimetres)
- bare_land_pct = numeric, proportion of site that is bare land (percentage)
- stoniness_pct = numeric, proportion of site that is stony (percentage)
- plant_litter_pct = numeric, proportion of site that is covered by plant litter (percentage)
- moss_cover_pct - numeric, proportion of site that is covered by moss (percentage)
- lichen_cover_pct = numeric, proportion of site that is covered by lichen (percentage)
- haylike_cover_pct = numeric, proportion of site that is covered by haylike vegetation (percentage)
- tree_cover_pct = numeric, proportion of site that is covered by trees (percentage)
- other_pct = numeric, proportion of site that is covered by other vegetation (percentage)
- vascular_agr = numeric, proportion of site that is covered by vascular vegetation (percentage)
- date = numeric, date of data collection
- time = numeric, time of data collection
- notes = character, miscellaneous notes from data collectors
- lon = numeric, longitudinal coordinate of site
- lat = numeric, latitudinal coordinate of site
- elev = numeric, elevation of site (metres)

Description of data processing

We were provided with 120 .xlsx files of the data from the 120 sites. We divided these equally between the two of us and then, using Google Sheets, copied and pasted the data from each file into one complete file comprising three tables (see above for details). Additionally, in Table 1, we added the value '0' to each empty cell in the 'cover' column and the value 'n/a' to each empty cell in the 'pheno_state' column. In Table 2, we added the value '0' to each empty cell in the 'cover' column. In Table 3, we transposed the data from columns into rows. This process took approximately 10 days.

After the dataset was completed and sent back to the commissioners, some basic analysis of the data was conducted and a number of visualisations created. These results and visualisations can be found below.

Discussion of results

The final result of the project is a dataset, where all the information from the separate data sheets is combined into three different tables. More comprehensive studies of the sites will be made later in the project, however to test the data and improve personal skills of the project group, some elementary visualizations were plotted (see Figures 3–6). In the following tables, species-richness values of study sites are plotted against soil temperature, elevation and soil freezing degree days.

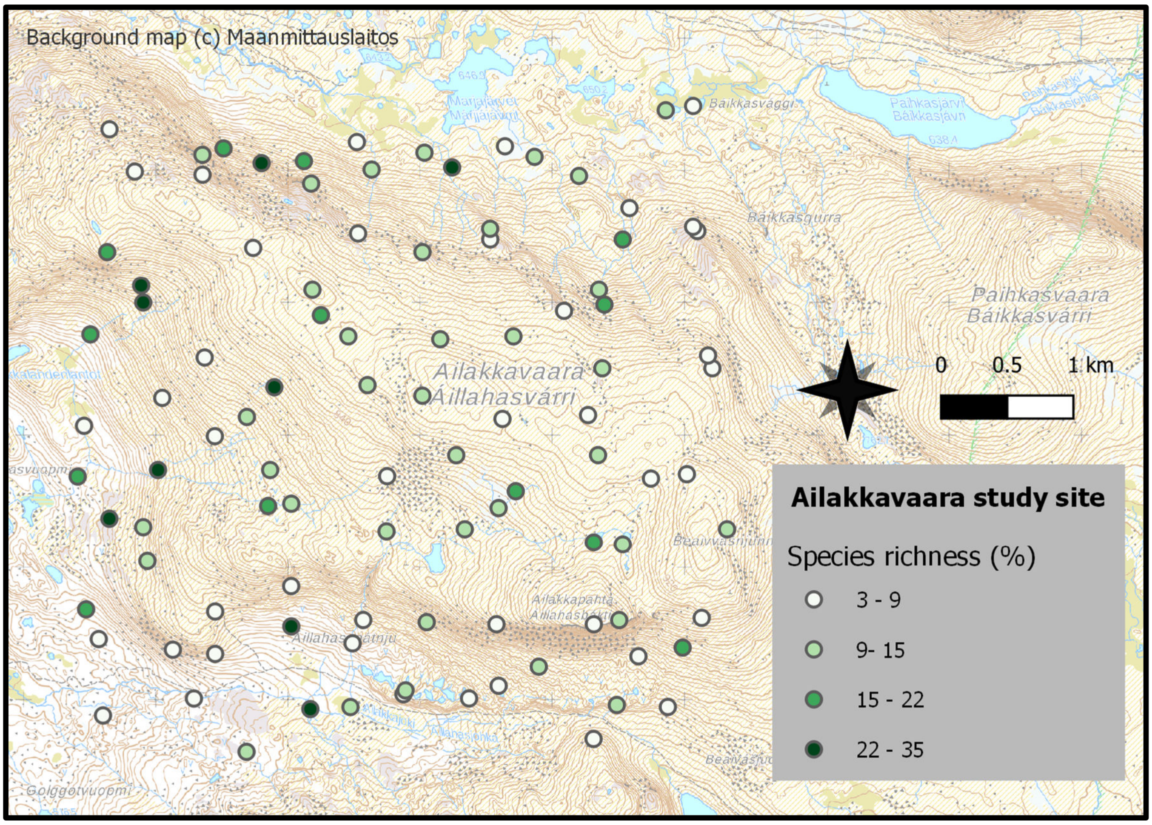


Figure 3. Values can also be examined spatially. Species richness of study sites at Ailakkavaara.

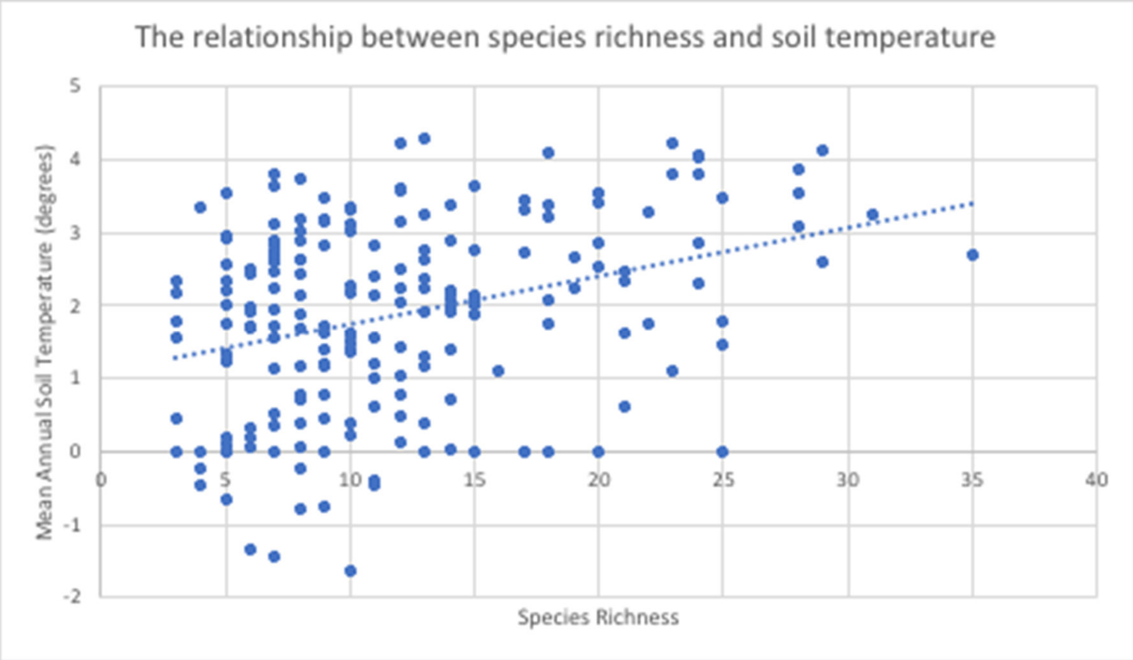


Figure 4. Species richness of study sites plotted against soil temperature.

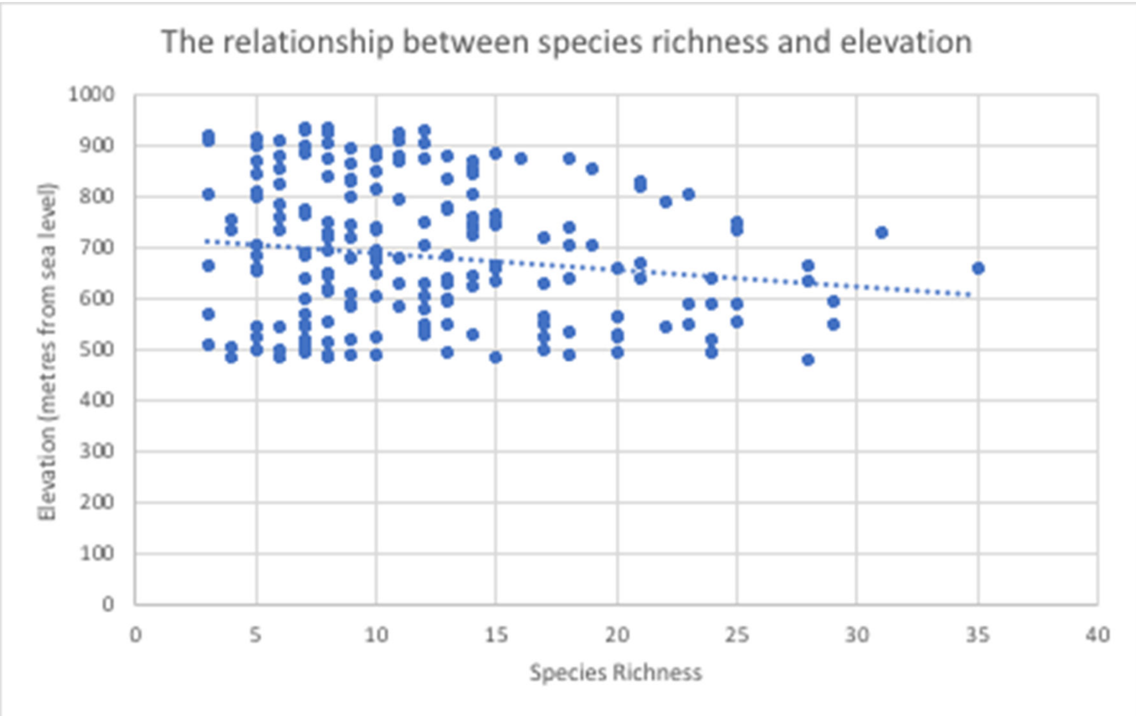


Figure 5. Species richness of study sites plotted against elevation.

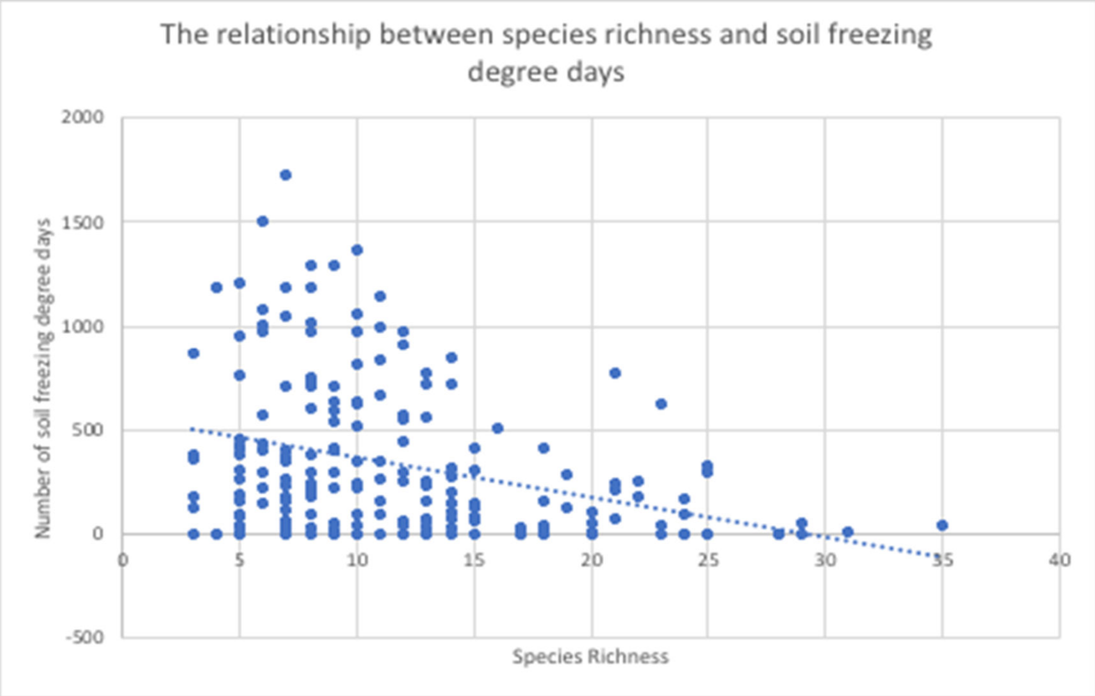


Figure 6. Species richness of study sites plotted against soil freezing degree days.

In this elementary analysis it was noted that there are few species present where the sum of soil freezing days is high. In contrast, species richness has a positive relationship with the mean annual soil temperatures, indicating there are more species present where soil temperature is warm. Results like these, with more careful analysis, could be utilized when attempting to understand how microclimate controls ecosystem functions over arctic and alpine environments.

Another possible use of the dataset is combining it with other existing information. In the project, species-richness values of the Ailakkavaara study site are studied against the DEM -Model provided by National Land Survey of Finland. The main emphasis of the project group was to visualize the site results on a map (see Figure 3), therefore it is wise to refrain from drawing conclusions based on the maps created during the project.

Conclusion

This was a relatively straightforward but important task, which provided a valuable opportunity to develop our excel and GIS skills. In addition, careful processing of large quantities of data was required. It is our hope that the final output will prove useful in this pioneering research project.

Perola, E. & Todorovic, S. (2020). Exploratory visual methods to aggregate origin-destination geodata. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 53–65. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter VI

Exploratory visual methods to aggregate origin-destination geodata

Perola, E., Todorovic, S., Muukkonen, P. & Järv, O.*

eero.perola@helsinki.fi, University of Helsinki
sara.todorovic@helsinki.fi, University of Helsinki
petteri.muukkonen@helsinki.fi, University of Helsinki
olle.jarv@helsinki.fi, University of Helsinki

**Corresponding author olle.jarv@helsinki.fi*

Abstract

This study used exploratory visual methods to aggregate Twitter mobility data in the Helsinki Metropolitan Area into postal code areas. Further, this study attempted to visualize the realistic movements of people in a custom-made network between the postal code areas. Twitter mobility data provided by the Digital Geography Lab at the University of Helsinki was used as the sample data. Study methods were also exploratory, comparing between two basic methods: custom networks and aggregation to polygons. Results were both visualizations of Twitter data and its users' movements in a postal code area level, and a base code for further development to fully visualize movement in the Helsinki Metropolitan Area.

Keywords: Movement data; Social media data; Visualization

Introduction

Movements and mobility patterns have recently gained a lot of attention in research (Andrienko & Andrienko, 2013). The emergence of big data suppliers, and particularly geotagged social media data has provided numerous possibilities to explore not only the places people visit, but also the frequency of their movements (Huang et al. 2015; Chua et al. 2016). Massive movement datasets are important and valuable to unveil mobility patterns of people in large metropolitan areas, and the knowledge is also helpful in transportation planning (Zang et al. 2013).

Movement data usually consists of object trajectories in space and time (Zang et al. 2013). In its simplest form, it has an origin point, destination point, and information about the time between those observations, and these trajectories can then be visualized. Visualizing movement flows in a way that describe the routes that people would use in reality is not simple, in fact, large datasets can create problems in visualization (Zang et al. 2013). Connecting two points across the city with a straight line as a *direct flow* does not depict the real movements of people, so the challenge has been to create more realistic movement flows. Only when the individual movements across space are aggregated into larger scales, such as neighbourhoods, different movement patterns can be detected (Batty, 2018). Therefore, visualization is an essential part of making sense of movement flows (Batty, 2018).

Recently there have been more and more attempts to model large movement datasets at aggregate level with different visual methods (Andrienko & Andrienko, 2013; Zang et al. 2013; Huang & Wong 2015; Chua et al. 2016; Batty, 2018; Graser, 2019). For example, Chua et al. (2016) visualized tourist flows in Italy using geotagged Twitter data and created a map with the movement flows and directions. More recently, Graser (2019) developed a Python library for analysing and visualizing movement data. These methods, however, require that the analysed data is in a cleaned form that suits the used methodologies, but even more importantly, user requirements include some more advanced programming skills and basic understanding of analysing big data.

This study is an exploratory analysis which partly builds on the work by the Digital Geography Lab (DGL) at the University of Helsinki. Recently, DGL has been exploring whether Twitter data can be used to estimate population density in Finland (Järv, 2020). The aim of this study is to explore methods to aggregate and visualize Twitter movement data into spatial units, and to create movement flows based on the frequency of movements. The ultimate goal of this study is to create a graph that takes a set of movements as an input, calculates the shortest route between the origin and destination, and updates the values of each edge that the shortest route passes through. Each edge would have a value depicting how many times a movement has passed between those two polygons. Methods in this study are exploratory and the topic is approached with a trial-by-error attitude. Given the novelty of the current methods for aggregating movement data, the restrictions in technical expertise of the authors, as well as other external issues during the Spring 2020, this article will describe the process of aggregating movement data as a learning experience. Hence, the study questions are:

1. What is the possible workflow to aggregate movement data to postal code polygons, and to create movement flows between the polygons?
2. What are the challenges and opportunities of visualizing movement data?

Data

Twitter data

Twitter user data was provided by the Digital Geography Lab (DGL) research group at the University of Helsinki. The complete dataset consists of geotagged Tweets in Finland from January 2016 to June 2019, collected from the Twitter API. This data has been pre-processed by the DGL, and it has information of the timestamp, a geotag (i.e. spatial coordinates), an anonymous user ID as well as the potential home country or municipality, calculated by the DGL (see Massinen, 2019).

For graph building, a sample dataset of 62K line records depicting individual movements of foreign visitors in the Helsinki Metropolitan (HMA) area was used. The movement line consists of origin and destination points in WGS84 coordinate system. In the data there is information of the timestamp at the origin point and the destination point, as well as the duration between the two points (i.e. Twitter posts) as days, hours, minutes, and seconds. Additionally, there is information about the distance or the length of the movement in kilometres between the points. The original data is visualized in Figure 1.

From the map one can see that the movements are densest in the downtown on Helsinki, and that other movement hotspots are for example Helsinki-Vantaa airport in the north of the map, as well as Suomenlinna island. Some of the movement lines are located at the sea, because probably some of the Tweets have been posted on the ferry between Helsinki and Tallinn.

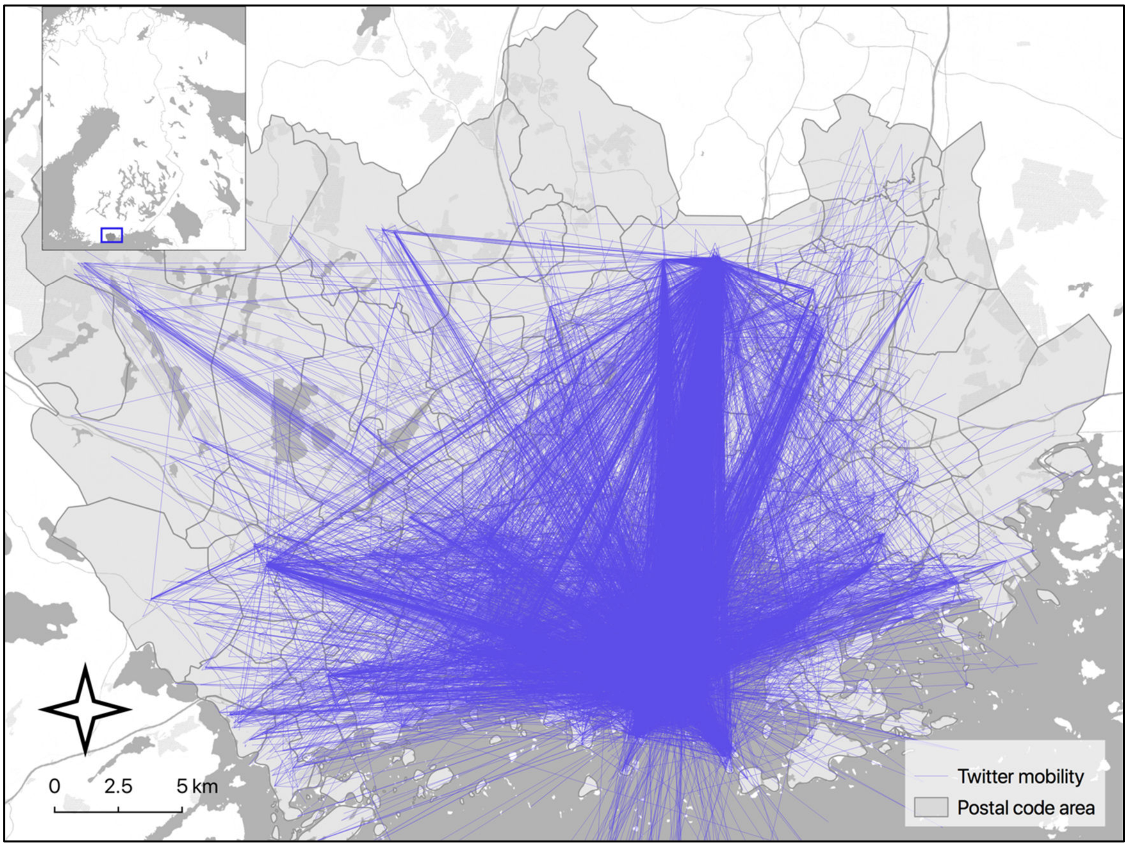


Figure 1. Original direct mobility patterns of Twitter users in Helsinki Metropolitan Area.

Other data

Paavo -postal code polygons for the Helsinki Metropolitan area were downloaded from the Statistics Finland WFS server (Statistics Finland, 2020). Each polygon holds also additional information about the demographic, socioeconomic and household structure, as well as information about jobs in that postal code area in 2017 and 2018. However, in this study only the polygon borders and the name of the postal area were relevant. In the HMA there are 168 postal code areas, and each one has a unique postal code number (e.g. 00100).

Methods

This study used exploratory methods to find out how movement flows could be visualized on postal code level. Analysis was done with custom made Python notebooks in JupyterLab, using open source libraries such as GeoPandas and networkX (Hagberg et al. 2008). Visualization was done in ArcMap and QGIS.

The first phase was to try to visualize the movements as flows between the postal codes in a network. A network graph that is applicable in network analysis in Python using networkX needs to consist of nodes and edges. Nodes, or vertices, are connected to each other by edges, and together they form a network.

Two different networks were tested and experimented to test different options to calculate routes between the origin and destination point of the tweets. The first option was to determine the fastest route through the OpenStreetMap network and to choose those polygons which this route traverses through. Here, the routing was done with the help of Henrikki Tenkanen's (2020) Python Routing Workshop. However, this option was soon found to be too heavy to automate.

Another method that was tested was to apply code from the Python Routing Workshop to a custom network and to calculate the fastest route through this network. A customized Triangulated Irregular Network (TIN) was created between all postal code polygon centroids in ArcMap. TIN is a 3D network that is used widely in GIS to represent surface morphology. TIN creates a triangulated network from a set of points, and results in a smooth network connecting the points. In order to get the nodes of the custom graph, the postal code polygons were converted into centroids. Custom editing was applied to remove impossible connections in the TIN network between some of the polygons, where it is not possible to travel by car or public transport. The connecting lines tried to imitate the possible routes in the Metropolitan area. For example, edges that would go through waterbodies with not existing bridges or ferry routes were removed, edges going through the airport or a big forest such as the Central Park or Nuuksio National Park were removed. As TIN is often associated with three-dimensional data and topography, the third dimension needed to be removed for network analysis. The network was converted from three-dimensional form to two-dimensional form and exported as an ESRI shapefile.

Next, the previous method was applied to the whole Twitter-dataset in three loops. First each tweet had a fastest route calculated between its endpoints. Here some of the tweets had identical start and endpoints, so those tweets were omitted. Also, some points (e.g. tweets from cruise ships) were not within the postal code polygons, so those tweets were again left out. As a result, a shapefile was exported with all the fastest routes within the area.

Further, the postal code polygons were iterated over within another loop iterating over the fastest routes. For each route, every polygon was tested if the polygon contained any part of the fastest route. If this was the case, an activity was added for the polygon in a column called ‘activity’. This number was then further updated with each route passing through the polygon adding to the activity-value.

Last, each start and endpoint were aggregated to the postal code polygons for easier visualizations of the original tweeting locations. With these data frames came the possibility of doing at least some kind of data exploration considering tweeting data and its users’ mobility throughout the HMA. In addition to these methods with concrete outputs, some exploratory methods were also tested, with mostly failing outcomes.

The workflow of the analysis is presented in Figure 2. Blue boxes indicate steps that did not work or could not be solved during the analysis.

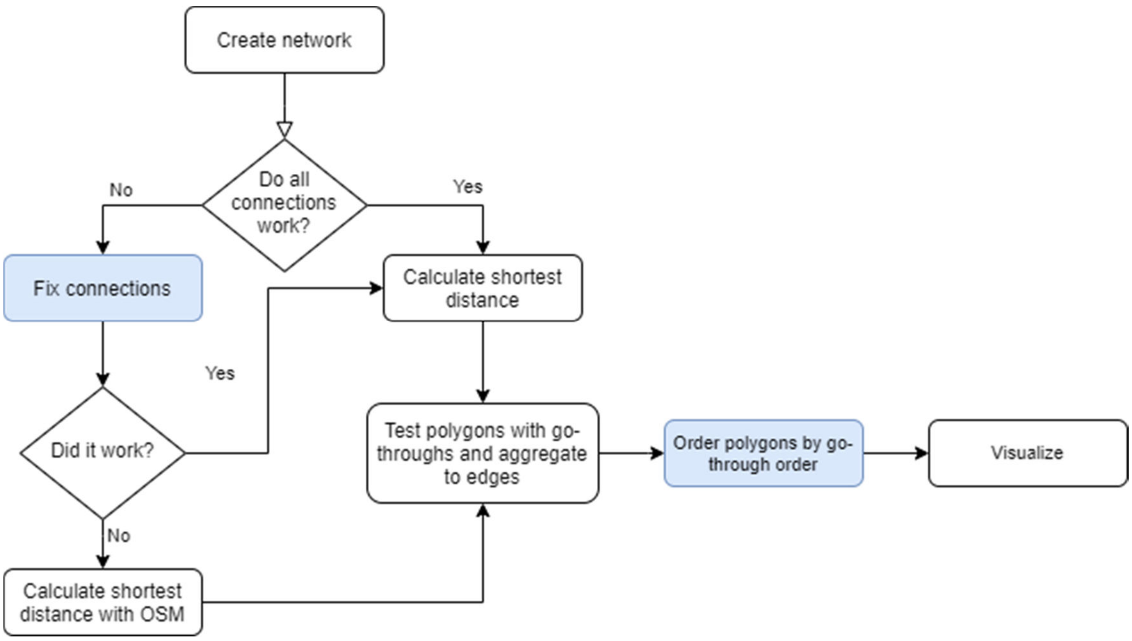


Figure 2. Workflow of the study. Blue boxes indicate steps that could not be solved.

Results

First the custom-made network was created. After cleaning up the original TIN, a network graph from the TIN shapefile was created by using the networkX library. As a result, a network where each edge stands for a possible route between postal code areas was constructed. Both the original TIN network after the removal of impossible connections, as well as the custom network graph are shown in Figure 3.

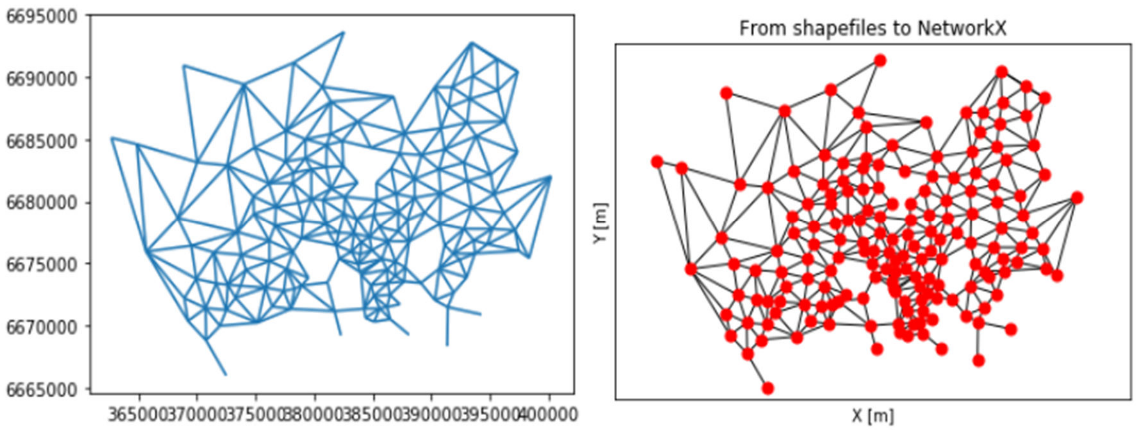


Figure 3. TIN network after removal of impossible connections (on the left) and the custom network graph created from TIN (on the right).

Next, a shortest path between two custom made points were calculated based on this network. As one can see from Figure 4, the resulting 'shortest path' was not the true shortest path in the custom network. If no weights are assigned in a networkX network, the shortest path is calculated by the edge lengths. Thus, the path calculated should not be like the one here.

Upon further investigation, it became apparent that when creating the network, the algorithm did not count all the edges as connections between its endpoints' nodes, thus resulting in failing path making. For an unknown reason some of the edges of the network were not counted as connections between its endpoints' nodes. This led to the routes calculated not being the actual shortest routes in the network between these two points.

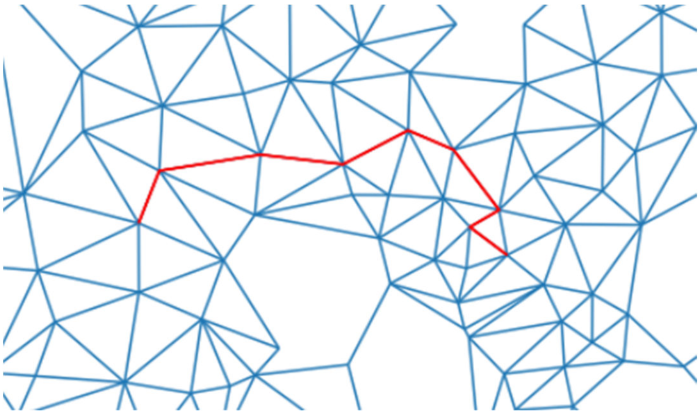


Figure 4. Example of a shortest path between one origin and destination point in the TIN network.

Further, the same approach was applied to a small subset from the original Twitter data. In Figure 5 is shown the original format of the data. When trying to create individual fastest routes through the network between e.g. a tweet with a starting point in the city center and an ending point at the airport, the routing tool could not find any connections between these relatively longer routes. This was due to networkX not creating a connection between all end nodes and their respective edges. This problem remained unsolved, and eventually led to abandoning the usage of the custom network for creating movement flows, and the focus shifted towards using Open Street Map to calculate the fastest routes as determining the tweeting activity in polygon areas.

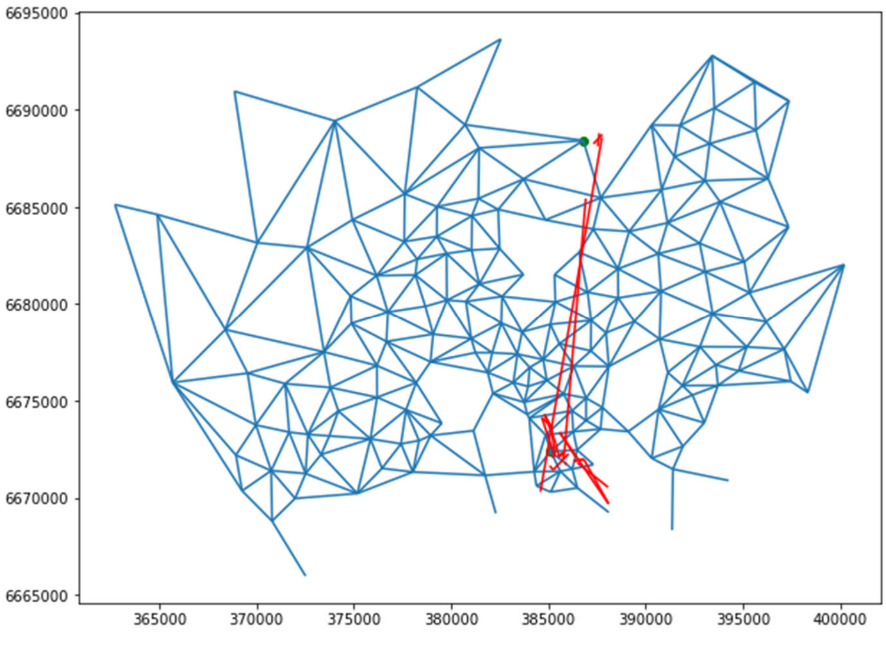


Figure 5. Example of original tweet route data.

Based on the fastest routes in the Open Street Map network between tweets, individual tweets were aggregated into the postal code areas to highlight areas where there is more tweeting activity. As is shown in Figure 6, most of the activity happens in the city centre of Helsinki, and a clear directional pattern can be seen in north-south direction. Tweeting is active also around the airport in the north of the study area. The graph in Figure 67shows the distribution of the tweets between different postal code areas. The city centre (Helsinki Keskusta – Etu-Töölö) accounts for 11 % of the tweets, followed by other central districs such as Kaartinkaupunki, Kruununhaka, Kallio, Töölö districs and then Veromiehenkylä near the Helsinki-Vantaa airport.

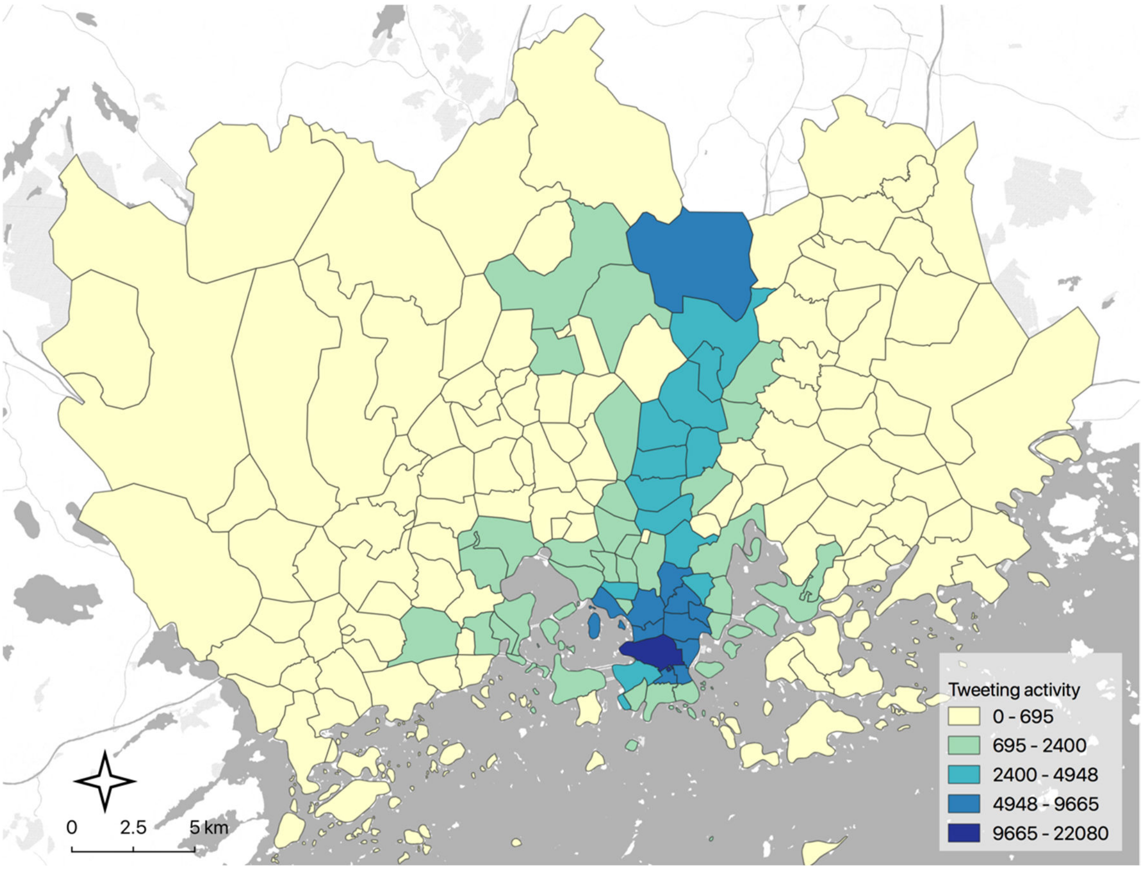


Figure 6. Individual Tweeting activity aggregated into postal codes in the Helsinki Metropolitan Area. Classification method is Natural Breaks.

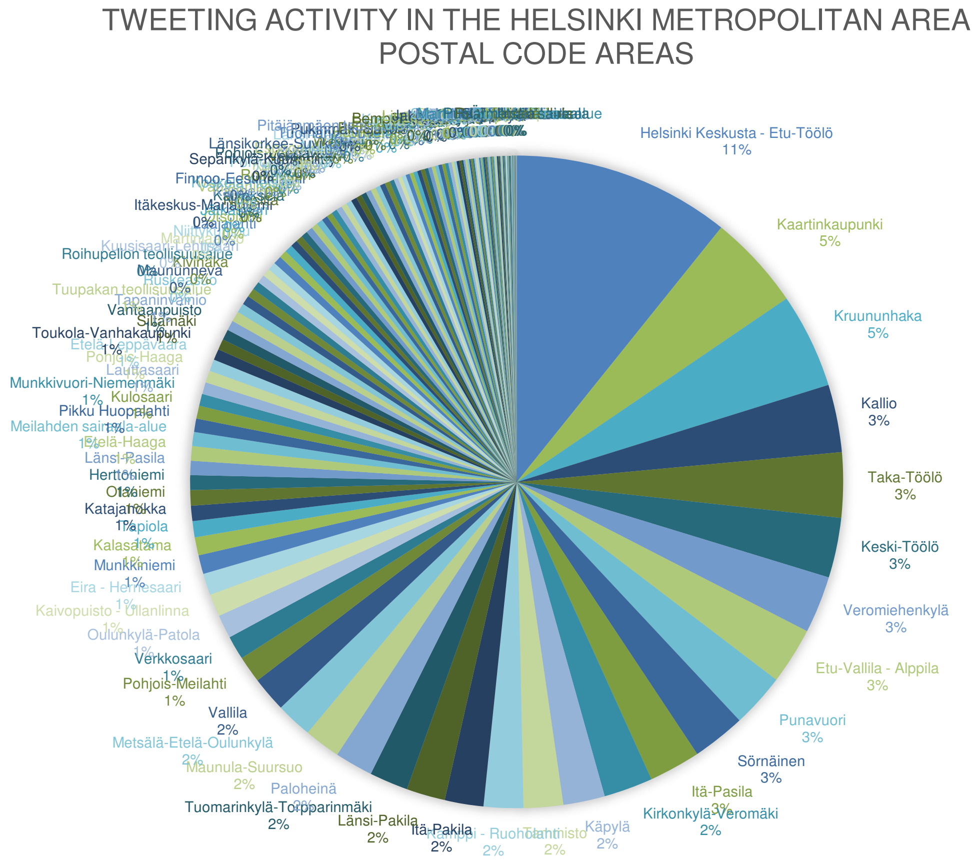


Figure 7. An illustrative graph of the distribution of Tweeting activity in different postal code areas in the Helsinki Metropolitan Area.

Discussion and conclusion

This study was an exploratory analysis, conducted in a trial-by-error manner, to find out how origin-destination movement data could be aggregated into bigger spatial units, and to create movement flows between the areas instead of straight lines between the points.

Aggregating individuals' tweets into postal code areas based on the fastest routes in the Open Street Map network between tweets revealed areas where the tweeting activity has been highest. This kind of visualization, although it is simple, it provides a picture of larger scale movement patterns (Batty, 2018). Areas in the city centre as well as around the airport are highlighted as the most active, where the city centre provides most of the services, events and other activities to tweet about, and the airport is highlighted possibly because it is an important place for tourists and travellers. While the choropleth map of activity frequency was only the first stage of revealing detailed movement patterns, further analysis is required to reveal more detailed patterns of the directions of those movements.

This study was ultimately unable to answer to the first study question, i.e. to create movement flows of origin-destination data on aggregate level. Several limitations and uncertainties were found during the process, emphasizing the requirements for this kind of analysis. First, the importance of detailed planning and required programming skills was highlighted during the analysis process. Many parts of the analysis could not be proceeded due to unknown errors and other dead ends.

Second, since the network analysis was ultimately made for the custom constructed TIN network connecting the postal code areas, there were several uncertainties regarding the results. First of all, the shortest path calculation does not have any weights, meaning that the resulting shortest path between the polygon centroids can be a bit different than it would in reality. Currently this network attempts to represent the real connectivity of different polygons, but on a very coarse resolution. Since the edges do not have information about the time that takes to pass the edge, with this network it is only possible to calculate the shortest route between edges. Even this is not remotely accurate, as all the edges do not count as connections between nodes.

Also, problems occurred when the custom-made example points for testing the shortest path were changed into the Twitter movements. While the example route looked somewhat as what was expected, the Twitter routes suddenly did not behave at all, even after multiple attempts with different subset sizes and rewriting the code several times. Thus, the attempt to create aggregated movement flows (such as in Chua et al. (2016)) with the shortest-path routing failed for a larger sample and has to be improved in the future attempts. Thus, next step would be to find a way to calculate the fastest route in the network. Furthermore, more detailed movement flows are also possible to visualize, for example, to include also the directions of the movements.

Ultimately, the code came close to creating aggregated movement flows of the HMA postal code areas. If the networkX code worked as planned, it would have created a data frame which would include activity data of movement flow through each connection between neighbouring areas with the help of the custom created network. Also, if a method was found to connect selected polygons in the right order to create routes between their central nodes, a solution to the initial study question 1 could have been accomplished. Thus, the next steps to follow from the code provided here would be to either:

1. Figure out how to add missing connections along edges to the custom network or
2. Figure out a method to connect polygon centroids with lines in a geographic order.

After one of these two would be achieved, one could be able to visualize these easily with standard methods.

Newly developed Python tools, such as Moving Pandas (Graser, 2019), provide useful methods to further explore origin-destination data and to visualize movements. Given the numerous improvements and opportunities to continue the development of the graph, the novel visual methods should be further exploited in the future and applied to the data used in this study.

References

Andrienko, N., & Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures. Information Visualization, 12(1), 3–24.

Batty, M. (2018). Visualizing aggregate movement in cities. Philosophical Transactions. Royal Society Publishing B, 373.

Chua, A., Servillo, L., Marcheggiani, E. & Vande Moere, A. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. Tourism Management,Volume 57, pp. 295-310,

Graser, A. (2019). MovingPandas: Efficient Structures for Movement Data in Python. GI_Forum – Journal of Geographic Information Science 2019, 1-2019, 54-68.

Hagberg, Aric A., Daniel A. Schult and Pieter J. Swart (2008). Exploring network structure, dynamics, and function using NetworkX. In: Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds). Proceedings of the 7th Python in Science Conference (SciPy2008). Pasadena, CA USA, pp. 11–15.

Huang, Q. & Wong, D. W. S. (2015). Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. Annals of the Association of American Geographers, 105:6, pp. 1179-1197.

Järv, O. (2020). Can we use Twitter data to estimate population distribution in Finland? Digital Geography Lab. Available at: https://blogs.helsinki.fi/digital-geography/2020/01/12/estimating-finnish-population-from-twitter-data/. Accessed on: 10.3.2020.

Massinen, S. (2019). Modeling Cross-Border Mobility Using Geotagged Twitter in the Greater Region of Luxembourg. (MSc thesis). University of Helsinki, Faculty of Science.

Statistics Finland (2020). Paavo – Postinumeroalueittain avoin tieto. Retrieved: 1.2.2020.

Tenkanen, H. (2020). Python Routing Workshop. Available at: https://github.com/HTenkanen/routing-workshop. Accessed on: 10.3.2020.

Zeng, W., Fu, C-W., Müller, S. & Qu, H. (2013). Visualizing Interchange Patterns in Massive Movement Data. Computer Graphics Forum, 32, 3.

Hirvonen, H., Leppämäki, T., Rinne, J., Muukkonen, P. & Fink, C. (2020). Modifying and analyzing Flickr data for wildlife conservation. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 66–90. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter VII

Modifying and analyzing Flickr data for wildlife conservation

Hirvonen, H., Leppämäki, T., Rinne, J., Muukkonen, P. & Fink, C.

hanna.hirvonen@helsinki.fi, *University of Helsinki*
tatu.leppamaki@helsinki.fi, *University of Helsinki*
jooel.rinne@helsinki.fi, *University of Helsinki*
petteri.muukkonen@helsinki.fi, *University of Helssinki*
christoph.fink@helsinki.fi, *University of Helsinki*

** Corresponding author* christoph.fink@helsinki.fi

Abstract

Applying social media data in researching protected area visitors can be useful in minimizing their impact on the biodiversity. An increased social media activity can be expected in national parks due to the growing nature-based tourism and the increasing use of social media. The tourism in national parks can lead to an impact on the area's biodiversity, resources, and environment. In this work, we study the possibilities Flickr data offers for conservation science, while aiming to provide methods for further research. We describe the dataset in multiple ways and examine the link between the accessibility and the frequency of social media posts. We create and utilize a script to merge geotagged social media point data with national park polygon data and global accessibility data, calculate social media post densities in national parks, and summarize them at the national and regional levels. We study point patterns in Sub-Saharan African national parks and create a kernel density raster layer of the Flickr posts in the region. Finally, we perform a cursory analysis of the linguistic content of the Flickr posts globally. Our results do not show a clear correlation between the Flickr post density in national parks and the accessibility from the nearest population center globally, which signifies a need for a regional examination or for a more sophisticated accessibility dataset. We find clear clusters of Flickr posts inside most Sub-Saharan national parks and have examples of national parks with concentrated and dispersed Flickr post distribution, although clustering is much more prevalent. Our linguistic analysis demonstrates the dominant role of English in Flickr, which might indicate an overrepresentation of people from English speaking countries in the data.

Keywords: Accessibility; Conservation; Flickr; Geoinformatics; Linguistic analysis; National park; Social media

Introduction

Social media data is useful in conservation science. Since the discipline tends to benefit from spatially precise data, it extracts social media data from different social media platforms nowadays such as Facebook, Twitter, Instagram, and Flickr (Di Minin et al., 2015). The data can be used in e.g. systematic conservation planning (Margules & Pressey, 2000; Knight, Cowling & Campbell, 2006) and in modeling species distributions (Elith et al., 2006). The data accuracy can be higher compared to more traditional data sources (University of Helsinki, 2015), the process more cost-efficient and continuous (Hausmann et al., 2017a), and the temporal and spatial resolutions better (Richards & Friess, 2015).

National parks and other protected areas have a great significance in wildlife conservation, protecting species, and possibly in reversing the biodiversity crisis (Watson et al., 2014). National parks are areas protected by the government to preserve the natural environment (Encyclopaedia Britannica, 2020). Conservation biology can be described as a mission-oriented discipline that aims to protect and restore biodiversity. It usually focuses on the issues that need quick action and could have significantly negative consequences. Because of the growing nature-based tourism (Balmford et al., 2015) and the increasing use of social media (Kaplan & Haenlein, 2010; Mayer-Schönberger & Cukier, 2013), increased social media activity in national parks can be expected. This leads to more available user content to further utilise in conservation research.

Tourism in national parks can be a double-edged sword. According to Di Minin et al. (2013), ecotourism has a potential to generate political support for protected areas. It can also generate funding for covering the park management costs (Buckley, 2009; Buckley, Morrison & Castley, 2016; Gössling, 1999) and has been promoted as a way to support biodiversity conservation and economic development (Goodwin, 1996; Krüger, 2005). Still, it can also lead to a detrimental anthropogenic impact on the biodiversity, the resources, and the environment of an area (Buckley, Morrison & Castley., 2016; Gössling, 2002). Particularly the biodiversity of small areas can suffer from the edge effect (Woodroffe & Ginsberg, 1998).

One of the top tourist destinations is Sub-Saharan Africa (World Tourism Organization, 2015). We choose to focus on Sub-Saharan Africa due to its density of well-known tourism-oriented national parks, which are typically safari parks (Africa Sun News, 2003; Crush, 1980; Siegfried, Benn & Gelderblom, 1998). Sub-Saharan Africa can be defined as the Africa south of the Sahara Desert, consisting of countries such as Ethiopia, Ghana, Kenya, Tanzania, Namibia, and Botswana. The African parks can support wildlife conservation while the potential use of social media data to inform conservation may increase in the future (Tenkanen et al., 2017; Willemen et al., 2015). The tourists are attracted to the African protected areas mainly by their charismatic megafauna. Hausmann et al (2017b) discuss the other important characteristics of nature-based tourism in Africa, the most essential of which are the biodiversity and the landscape aesthetics. Also, particularly when studying tourism in these areas, geographical factors, such as accessibility and human influence, can be very important. Consequently, it can be deduced that accessibility can be utilised in conservation science.

The indicators of accessibility often are different distance measures and travel times (Frank et al., 2008; Mavoa et al., 2012). It can also have a great effect on the post activity and the number of park visitors. Hausmann et al concluded in their study (2017b) that accessibility was a strong predictor of the user and the post activities, meaning that accessible areas tend to have more social media posts and active users. The study also revealed that the richness of charismatic species did not influence the social media use in the protected areas of Africa but rather of importance were the socio-economic conditions of the countries and their geographical characteristics.

Hausmann et al (2017) also note that the biodiversity and the environment of accessible areas can be threatened by a high human pressure. The disturbance on the area's biodiversity can include stamping down the vegetation (Pickering & Hill, 2007), disrupting the feeding and breeding of the fauna (Bouton et al., 2009; Ranaweerage, Ranjeewa & Sugimoto, 2015), and decreasing the successful reproduction (Steven, Pickering & Castley, 2011). Overall, the sustainability of nature-based tourism is indeed challenged (Buckley, 2011).

Tenkanen et al (2017) state that evaluating the benefits of the recreational value of national parks is often a crucial part of justifying the existence of these parks, which creates a firmer base for maintaining these areas for biodiversity conservation. Monitoring the visitor rates from social media data can be used in assessing the area's recreational value. Based on this it is possible to conclude that applying social media data in researching protected area visitors can serve as a justification for conservation.

Social media data usually contains text, images, videos, and tags. When using the data for e.g. conservation studies, it can be restricted by search parameters, such as keywords (Di Minin et al., 2015). The posts also contain the time stamp and possibly the location data. Because of these features, social media data has uniquely great spatial and temporal resolutions of populations (Longley, Adnan & Lansley, 2015), which makes it a very suitable data source for conservation science, although the use is still limited (Di Minin et al., 2015). To access the content, ready-made application programming interfaces can be used (University of Helsinki, 2015), and in publishing the data, the user privacy has to be taken into account.

The social media posts can tell about the preferences and the engagement of the national park visitors (Hausmann et al., 2017a; Levin, Kark & Crandall, 2015; Su et al., 2016). The data can be useful in the national park management as information for minimising the impact of the visitors on the area's biodiversity (Cessford & Muhar, 2003) and for understanding the interests of the visitors for promotional purposes (Hausmann et al., 2016), along with marketing purposes (Buckley, 2009; Smith, Verissimo & Macmillan, 2010; Tenkanen et al., 2017). It may be profitable for the park management to use data from these kinds of novel sources instead of carrying out the surveys themselves, which can be comparatively time-consuming and costly (Hausmann et al., 2017a). For example, the data might reveal the species the visitors have spotted or their favourite species and landscapes (University of Helsinki, 2015). There are still some weaknesses in studying social media data from national parks. For instance, the data tends to perform better in the parks with more visitors, and sometimes the visitor statistics and the user activity do not match (Tenkanen et al., 2017). Social media data also tends to be biased to the developed countries (Di Minin et al., 2015).

Established in 2004, Flickr is among the oldest social media platforms. It has some good qualities as a data source for conservation science, which is one of the main reasons we use data mined from it in this study. The site is popular among photographers and is commonly used for image sharing. The study done by Hausmann et al (2017) had results on the features of the Flickr users that were visitors in protected areas. They were described as experienced tourists and nature enthusiasts with interests towards some of the less charismatic species. In South Africa, it had the highest correlation with the official statistics. (Tenkanen et al., 2017.)

In this article, we study multiple aspects of Flickr posts in national parks. First, we inspect the relationship between the post frequency and the accessibility of a park in different spatial scales. Then we study the patterns the posts create within the parks of a chosen subregion, Sub-Saharan Africa, and look at example parks to understand where and why the posts are clustered. We focus particularly on Flickr data from Serengeti National Park and Nairobi National Park. Serengeti is a famous area of 14,763 square kilometres in Tanzania and Kenya that attracts visitors with its rich natural resources, mainly biodiversity and its highest large mammal density of the world (Eagles & Wade, 2006; Serengeti National Park). Nairobi National Park is a smaller park in Nairobi, the capital city of Kenya. Finally, we do a tentative inspection of the linguistic content of the posts. All of our research steps aim to explore the dataset and create methods, thus assisting the future research in studying accessibility and Flickr data in conservation science.

Data and methods

We employed three datasets to examine global accessibility, national parks, and social media posts. These are, respectively, the global accessibility to cities by the Malaria Atlas Project (Weiss et al., 2018), the World Database on Protected Areas (UNEP-WCMC & IUCN, 2020), and a dataset of Flickr posts represented as coordinate points. Accessibility is, in the global raster surface by Weiss et al. (2018), defined as the travel time in minutes from one raster cell to the nearest urban centre. Urban centres are areas with a high population density or a high number of built plots coinciding with at least 50,000 inhabitants. Travel time is quantified by measuring the combined effects of different highways, land features, and national borders. The data dates to the year 2015, and its spatial resolution is 30 arc seconds, or roughly 1 km² at the equator (Weiss et al., 2018).

The World Database on Protected Areas (WDPA) is a global collection of land and sea areas that hold high natural or cultural values and meet the standards for a protected area (UNEP-WCMC, 2019, pp. 8–9). The data consists of polygon boundaries of the protected areas, provided by various governmental and other entities. We utilised a subsection (n=2556) of the data, the areas labelled as 'natural parks'. Finally, we used Flickr posts that are geotagged coordinate points falling within the natural parks. After filtering the data for exact duplicates, we were left with over 2.3 million posts with a temporal extent from the year 2004 to January 2019. Attributes of the posts, such as the title, the textual description of the image contents, the accuracy of the positioning, and the URL of the photo were included alongside the location.

For our research, we utilised various open source geospatial software, mainly Python and QGIS (see the whole workflow in Appendix A). We began by filtering both WDPA and Flickr datasets: the first for entities marked as 'national park', and the second for duplicate posts. If a park consisted of different zones, we interpreted them as parts of the same park. Parks that lacked Flickr posts altogether were dropped. In addition, some of the national parks had overlapping boundaries which means that some Flickr posts fell within more than one national park. The exact number of these posts was 14,067. We used these datasets to calculate the average post density in each park. Density is defined here as posts per square kilometre. The areas of the parks were included in the WDPA dataset and included marine regions. We used the 'GIS_AREA' field as the indicator for the area extent. The densities were also summarised at the national and regional levels. The summaries were conducted by adding the number of the region's Flickr posts together and dividing it by the summarised area of the region. We also included some simple descriptive statistics for each level.

National parks were also combined with the global accessibility raster dataset. Different statistical values for each park were calculated by defining the park borders as zones and summarising all raster cells that fall within. This produced for example the minimum value it takes to reach the park in minutes, that is to say, how accessible the park is in the best-case scenario. The accessibility dataset was then joined with the Flickr post density dataset. Some parks are on islands where the accessibility dataset does not reach, so those parks were dropped from the accessibility statistics but were still included in the post density statistics. As with the post density statistics, accessibility statistics were summarised at the national and regional levels. We then tested whether correlation exists between the park accessibility and the post activity: the assumption being, that an easier accessibility would lead to more visitations and therefore to an increased social media posting. We used the minimum value of accessibility for each park, since the parks can be large, and some sections especially outside the road networks can be highly inaccessible. We assumed the minimum value within a park might be its entrance, since it is connected to a larger road network and thus captures the park's accessibility for the average visitor. Pearson correlation coefficient and scatter plotting were done for the minimum value and the post density both globally and regionally.

We then focused on Sub-Saharan African national parks and the concentration of the Flickr posts within them. Three methods to study the point patterns and their dispersion were employed: the goodness-of-fit test based on the quadrat counts, the Ripley's K Function, and the kernel density estimation (*KDE*). To increase the reliability of the point pattern analysis, we limited our scope to the parks with at least 500 posts. The first method simply tested against the null hypothesis that the points are randomly distributed across the study area by applying a uniform grid across the area and counting the points in each cell, or quadrat (Anselin, 2015). Then the probability of the pattern being random was tested with the Pearson χ^2 test, the alternative hypothesis being spatial clustering. After that requirement was satisfied, clustering and dispersion were tested in different scales using Ripley's K Function (Gillan & Gonzalez, 2012).

In the final approach, we created a kernel density raster layer of the Flickr posts in the region using the QGIS Heatmap Plugin with a quartic kernel shape, a grid cell size of 1x1 km, and a radius of 10 km. Universal guidelines for parameter selection appear sparse with case-by-case evaluations for each dataset being more common. Harth and Zandbergen (2014) propose that the grid cell size has little effect on the predictive accuracy. However, too large of a cell size creates coarser results but a smaller grid cell size can increase the processing time of the algorithm. Both Hart and Zandbergen (2014) and Garcia et al. (2015) note the importance of the bandwidth, or the radius, on the final results. We chose the radius based on the size of the national parks and the fact that Flickr posts seemed to be fairly concentrated in general. A larger radius would have saturated the highly concentrated areas, and a smaller radius would not have provided enough distinction between the areas when the pixel size is taken into consideration. We then chose two example national parks, one with a high concentration and one with a low concentration of Flickr photos and created kernel density maps of them. Because of the small size of the

Nairobi National Park, the example park with a low concentration of Flickr photos, we decided to use a grid cell size of 100x100 m and a radius of 1 km for its kernel density map, retaining the cell size to the radius ratio.

Finally, we studied the linguistic content of the Flickr posts by determining the likely language used and examining the most common words in the posts. The text was first pre-processed by discarding various non-linguistic features, such as URLs. Also, to get more accurate results in the language detection, the minimum length of the texts was set to be 15 characters. We noticed the Flickr users describe the images to varying degrees in both the title and the separate description field. Because of this, only posts where both these fields meet the requirements were used. Filtering left a total of 227,434 posts. The language of each post was determined using the Python implementation of the Langdetect library (Shuyo, 2010) which supports the detection of 55 languages. Short texts and multilingual posts created uncertainty in the detection, which is why a post was deemed identified only if the confidence of detection given by the software is over 85 %. Lastly, the posts were filtered for stopwords that occur often but have low semantic information, such as *the* or *his*. The most important words of the dataset were determined by a simple word count and by the term frequency-inverse document frequency (*tf-idf*) method. Tf-idf was used to highlight the terms that are frequent in single documents (in this case, posts) but not in the whole dataset, like the aforementioned stopwords. It is a widely used method in information retrieval and variations of it have been utilised in e.g. summarising recent events on Twitter (Alsaedi, Burnap & Rana, 2016). We employed it to attempt to highlight the infrequent words that still summarise the common topics discussed in the data.

Results

Accessibility and post density

Our results on Flickr post densities in national parks show that post densities are the highest in small parks. An interesting result is that of the top 15 national parks with the highest Flickr post density, five were in the British Virgin Islands. Another noticeable statistic is that many Israeli national parks rank high on the list as well. The highest number of Flickr images was in Yosemite national park in the United States. Overall, 12 of the top 14 national parks with the highest number of posts are in the United States. On the other hand, the parks with the lowest Flickr post density, all have an area over 10,000 km² and were in Canada, South Sudan or Venezuela. At the national level both British Virgin Islands and American Virgin Islands rank on the top in regards of the post density. Overall, island nations rank high on the list. Map of the densities at the national level is presented in Figure 1. Most of the nations with the lowest Flickr post densities are in Sub-Saharan Africa. The highest density region is Eastern Asia with an average of 8.55 Flickr posts per km² (Figure 2, Table 1). The highest number of Flickr images has been posted in Northern America. Excluding Greenland, the lowest densities are in Central Asia and Northern Africa.

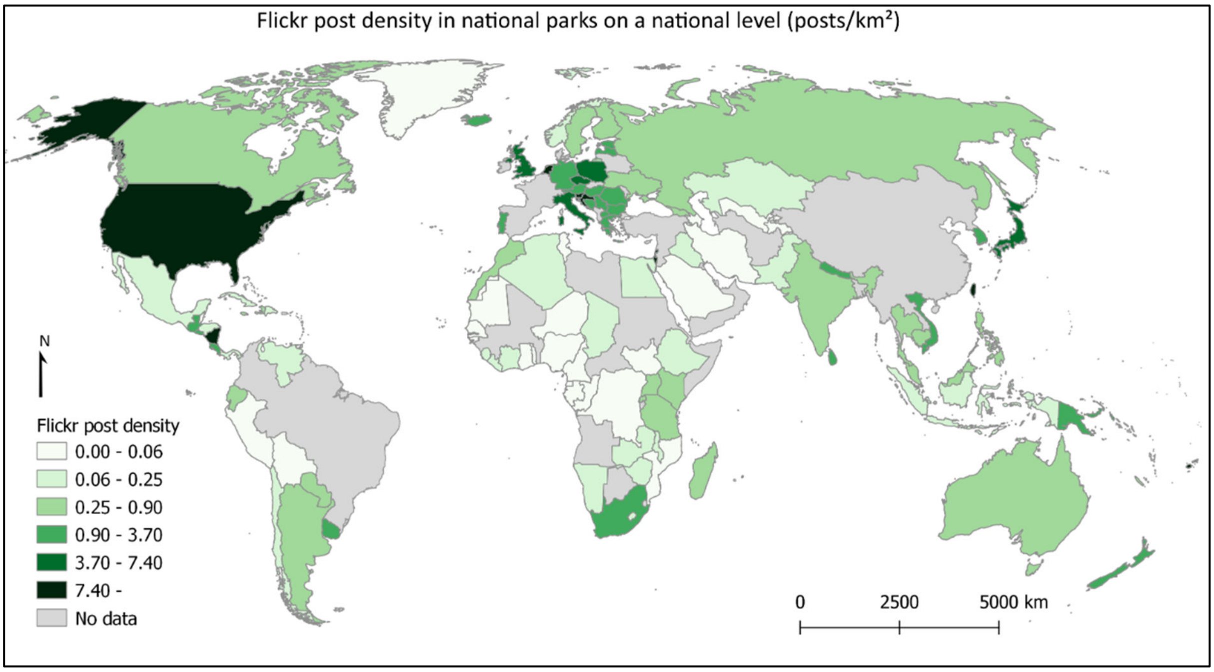


Figure 1. The Flickr posts per km² in national parks at the national level.

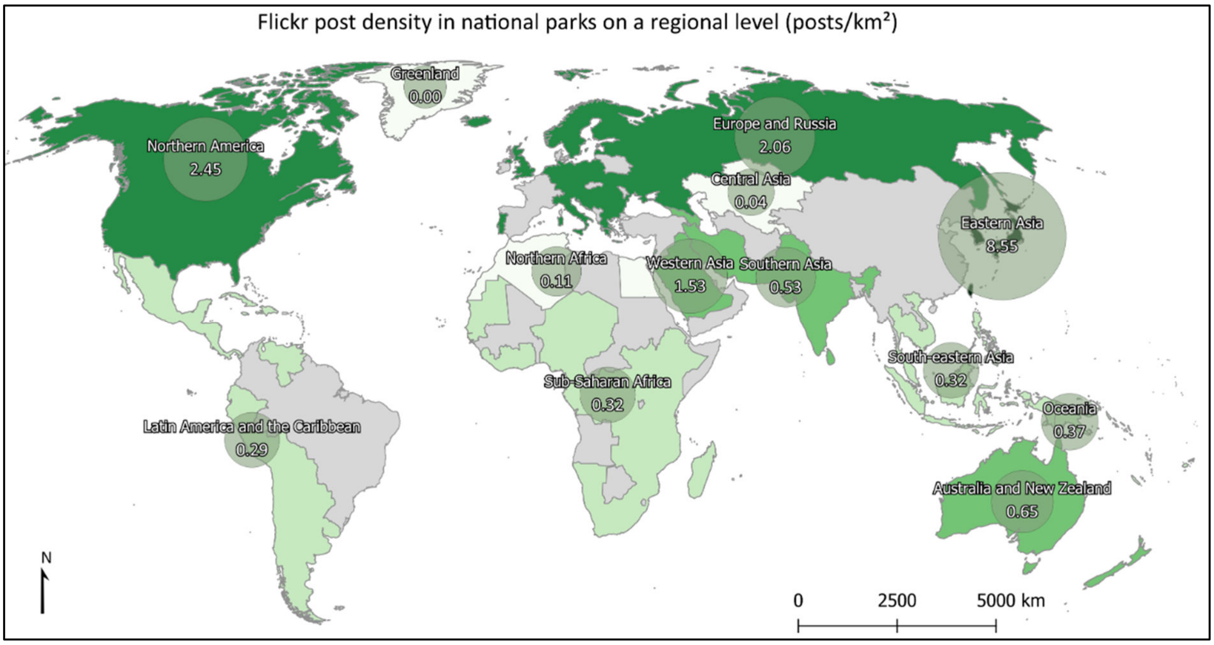


Figure 2. The Flickr posts per km² in national parks at the regional level.

Of the national parks in our dataset that have at least one Flickr post within their area, 125 have a minimum accessibility of 0 minutes. The Flickr post densities of these parks vary from 0 to over 8,500, which is the highest density of any park with accessibility data available. The average minimum accessibility of all the national parks is 162.73 minutes. At the national level, Netherlands has national parks with the highest median minimum accessibility (the lowest number) when the nations with at least five national parks are included. Within these countries, Canada is on the other end of the list with its lowest median minimum accessibility. At the regional level, Eastern Asia and Western Asia have the national parks with the highest median minimum accessibility, while Northern America has the lowest accessibility (Table 1).

Table 1. The summary statistics of the Flickr post densities and the national park accessibilities at the regional level. In the "Number of Parks" column, the number inside the parentheses indicates the number of parks that are included in calculating the park accessibility indices.

Region	Number of parks	Park area total (km²)	Number of Flickr -post total	**Post density (posts/km²)**	Accessibility index (median)
Eastern Asia	41	24,032	205,477	**8.55**	**13.00**
Northern America	99	426,094	**1,044,805**	2.45	123.00
Europe	338 (336)	198,868	410,160	2.06	43.50
Western Asia	83 (81)	15,185	23,292	1.53	16.00
Australia and New Zealand	**500 (497)**	309,351	201,723	0.65	115.00
Southern Asia	89	72,137	38,060	0.53	27.00
Oceania	13	11,024	4,025	0.37	41.00
Sub-Saharan Africa	205 (203)	533,901	170,699	0.32	66.00
South-eastern Asia	194	182,918	58,239	0.32	44.50
Latin America and the Caribbean	350 (339)	607,915	179,121	0.29	63.00
Northern Africa	25	93,183	10,068	0.11	49.00
Central Asia	4	30,685	1,258	0.04	62.50
Greenland	1	**961,673**	689	0.00	2,320.00

According to our analysis, there is no clear correlation between the density of Flickr photos and the minimum value of the accessibility index between the national parks (see Figure 3). The variation between the parks is large, and there are a great number of outliers on both axes: the parks that have either an extremely high post density or a poor accessibility. These parks lie on all regions without a clear pattern. Aggregating the values to the regional level (Figure 4) reveals a weak negative correlation (the post density increases as the time to access the parks decreases), but the data points are too scattered to call this result robust.

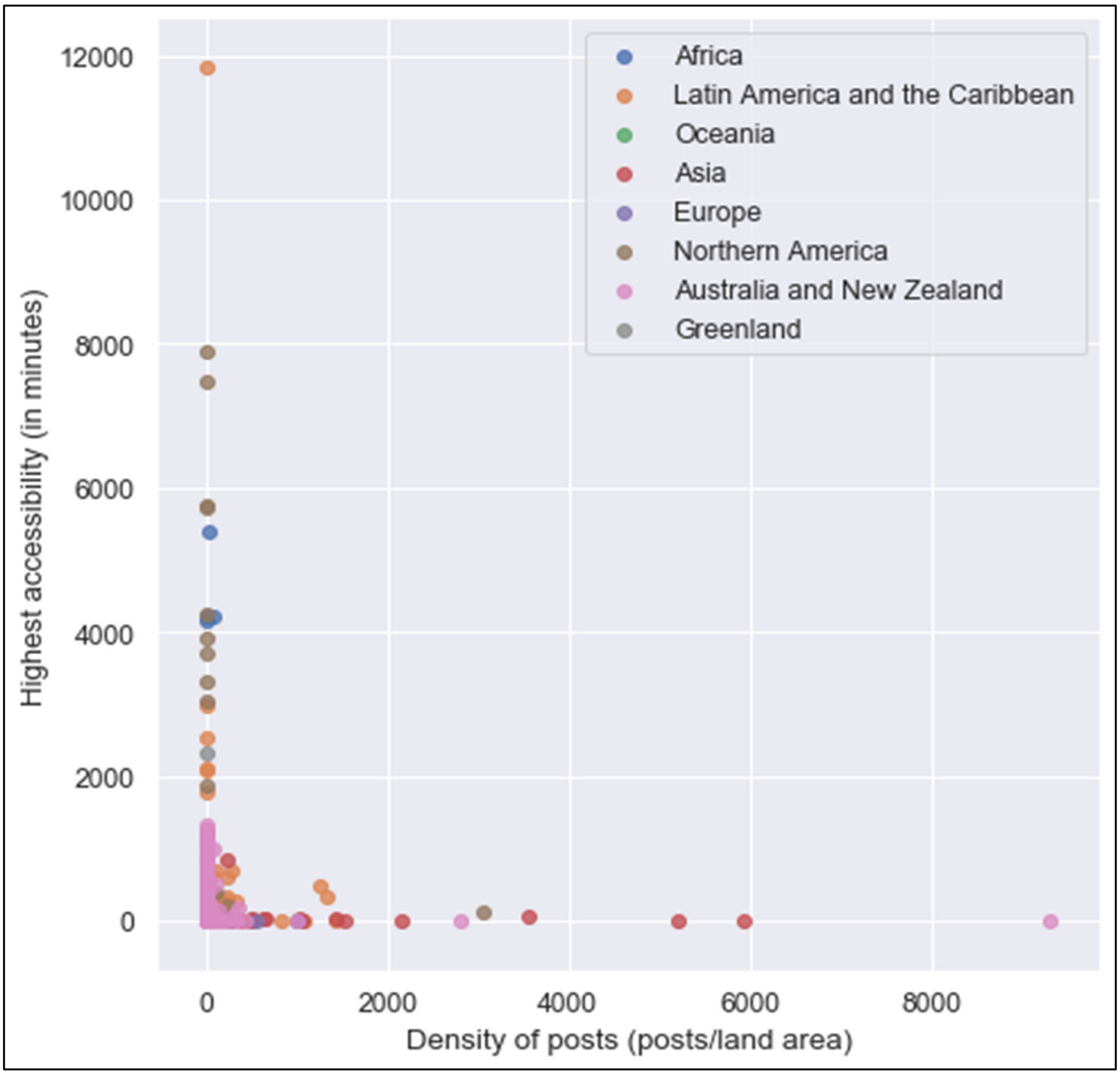


Figure 3. The most accessible (minimum value) section of the parks, and the density of Flickr posts per km². The region park *n* is in is marked by the color scheme.

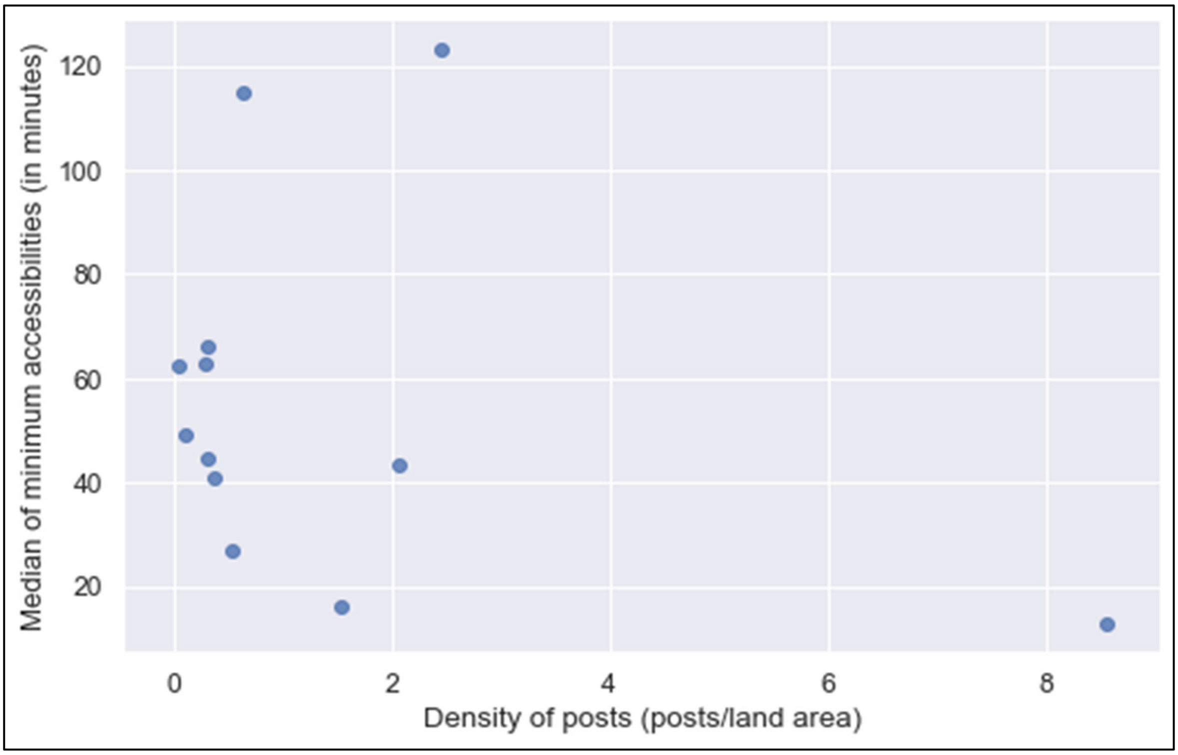


Figure 4. The accessibility and the post densities aggregated to the regional level. The Y-axis is the median of all the minimum accessibility values in the region. The X-axis is the total amount of the posts made in the region divided by the total amount of land area.

Post dispersion in parks

We first analyse the post dispersion within the parks in Sub-Saharan Africa using the quadrat count method. The assumption of the random dispersion could be rejected with a high confidence for all the parks analysed. The analysis of the Ripley's K results suggest clustering patterns for nearly all the parks at multiple distances (see Appendix B e.g. from Serengeti). Using this knowledge and visual information, we select two candidate parks for further inspection. The two resulting heatmaps from the kernel density estimation are illustrated in Figure 5 and Figure 6. Serengeti National Park (Figure 5) is an example of a park with a high concentration of Flickr posts. The concentration is high especially along the primary road which intersects the park in the northwest-southeast direction. The highest concentration of posts is in the middle of the park where the Serengeti Visitors Centre is located. The kernel density map of Nairobi national park presents a more dispersed example of Flickr posts within a park. It is much smaller than Serengeti and in the Kenyan capital city of Nairobi. The only part of the park with a slightly higher concentration of Flickr posts is located near the west-side border of the park where the Sheldrick Elephant & Rhino Orphanage is located. The park is filled with small roads which might explain the low concentration of the Flickr posts.

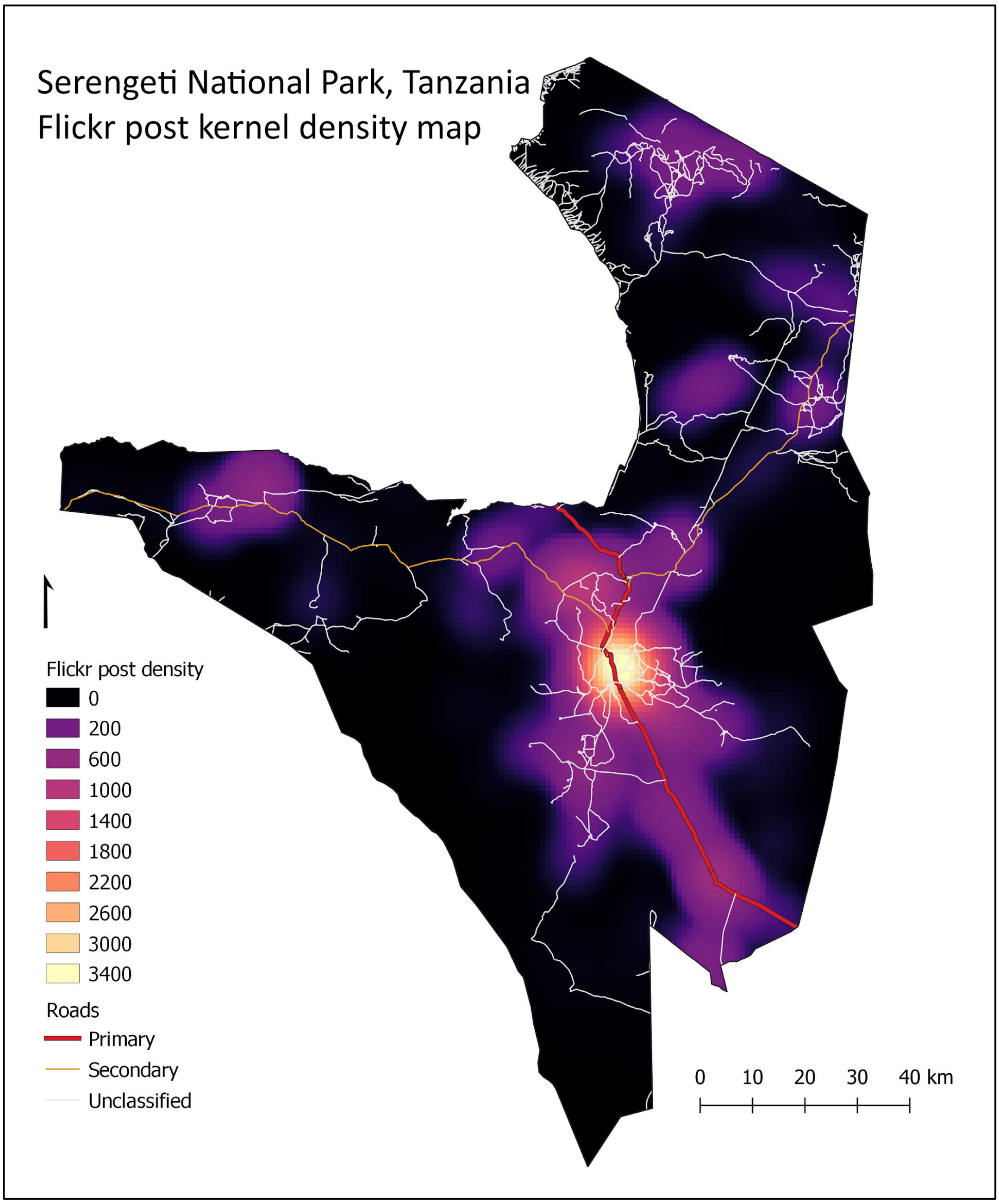


Figure 5. The Flickr post kernel density map of Serengeti National Park, Tanzania. The overall Flickr post density of the park is 1.30 posts/km². Road network © OpenStreetMap contributors.

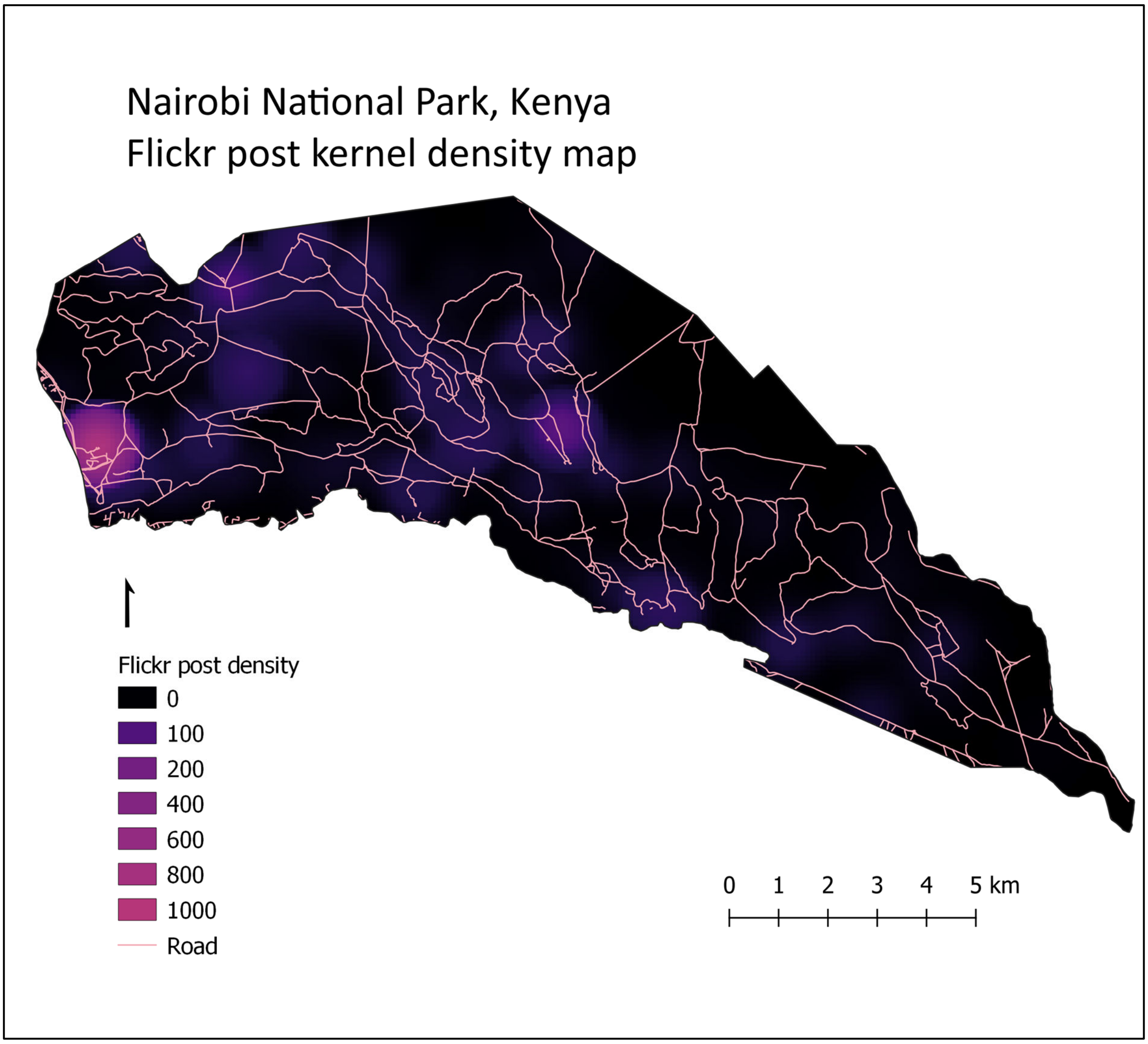


Figure 6. The Flickr post kernel density map of Nairobi National Park, Kenya. The overall Flickr post density of the park is 17.41 posts/km². Road network © OpenStreetMap contributors.

Linguistic analysis

The linguistic analysis of the posts that meet the criteria defined in the methods section shows that an overwhelming majority, about 80 % of the posts, were made in English (Figure 7). The second largest group is too uncertain to be identified, perhaps due to multilingual posts. Rest of the languages, including major world languages such as Spanish, make up only about 10 % of the posts. To learn more about the largest subsection of the posts, we look at the most common and important words in the posts identified as

English (Table 2). Simple word count highlights descriptive words of the surroundings, like *lake*, *mountain,* and *valley* as well as different US national parks (*Grand Canyon, Yellowstone*). Note that the count is case-insensitive and a word like *lake* could be a part of a place name. A tf-idf column does not share any entries with the word count, but similarly highlights the nature aspect of these parks – the words include plant and animal species (*pochard*, *acutifolia*).

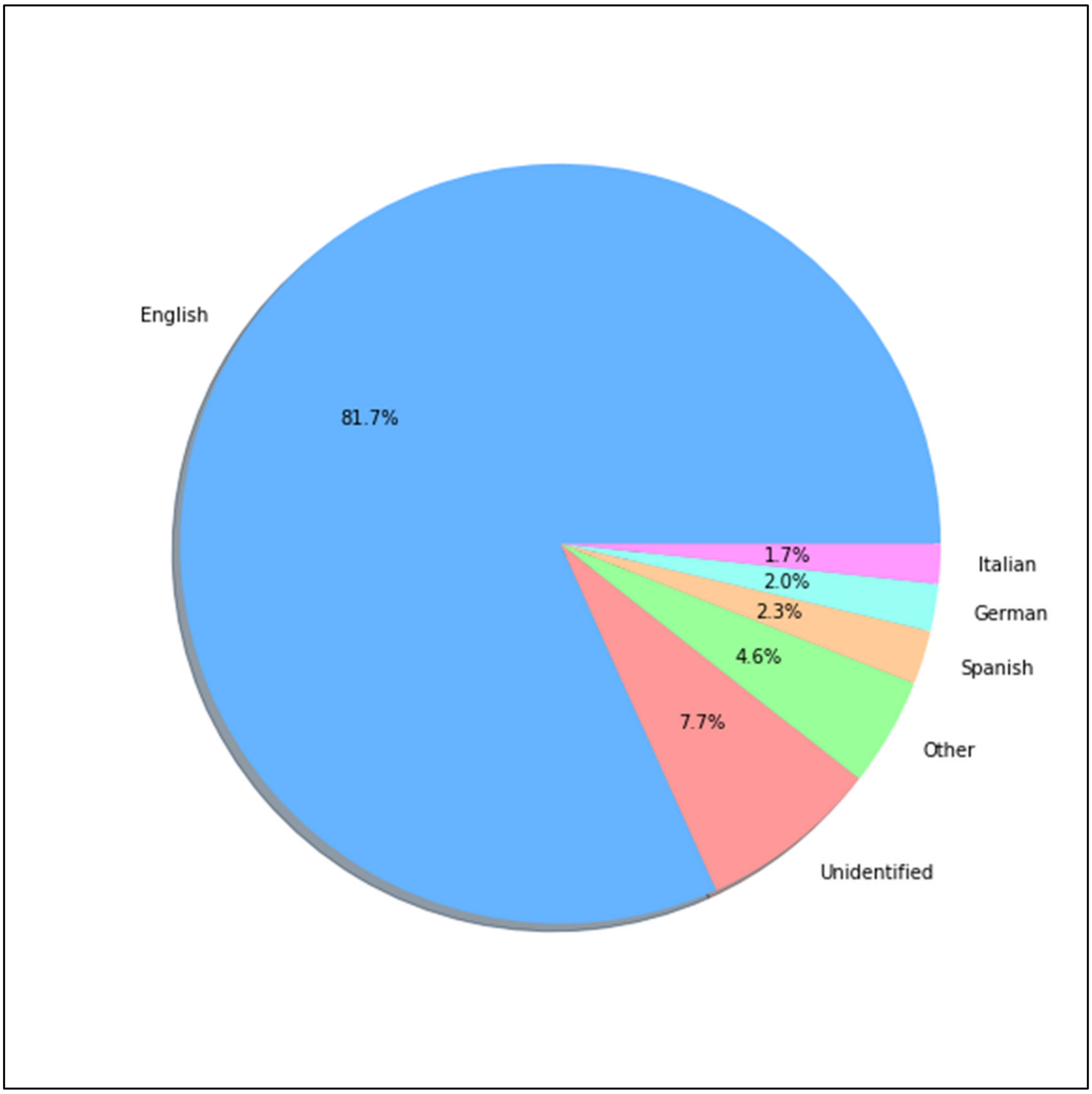


Figure 7. The proportion of the languages identified in the posts.

Table 2. The top 15 words of interest in the English language Flickr posts by the word count and the tf-idf methods.

Word count	***Tf-idf***
park	*pochard*
national	*therates*
canyon	*corunastylis*
lake	*troglodytes*
trail	*sochi*
view	*albiventris*
one	*arabian*
river	*jabba*
south	*queulat*
valley	*culprit*
grand	*acutifolia*
mountain	*tartally*
yellowstone	*flowerpecker*
yosemite	*whitehorn*
day	*abys*

Discussion and Conclusions

The Flickr post density seems to depend on the popularity of the application. The three regions with the highest number of Flickr posts (Northern America, Eastern Asia, and Europe) also have the highest post density within their national parks. However, when examining the national parks inside a country or a region, the results can be comparable. For example, the differences in the Flickr post densities between Canada's remote and large national parks and Yosemite national park in the United States can be considered representative of the differences in the visitation rates. However, additional information, for example, from other sources of digital data or information about the population and the tourism in the areas, should be included for more in depth conclusion of the actual rates.

No notable correlations between the accessibility and the post density of the parks are found either at the global or regionally aggregated scale. Multiple reasons can explain this. First, the national park dataset is heterogeneous: the parks are of widely different scales, some consist mostly of water, some have next to none posts, and some are on remote islands. The term 'national park' thus encompasses so many different types of areas that studying them together at the global scale proved difficult. The WDPA dataset does not seem to be exhaustive either, as, for example, there are no parks marked in e.g. France, Spain and China. Further research might benefit from studying only a clearly defined subset of the parks, such as parks of predefined scale, in a single country, or with a minimum number of posts in them.

Second, the accessibility dataset by Weiss et. al (2018) has some shortcomings when used in this manner. It measures the accessibility to major population centres which might not represent how accessible the park is to its visitors. For example, the population in the surrounding areas might be dispersed enough not to qualify as a population centre. The parks that are popular with tourists might also be more accessible due to regular transportation connections. It is also possible that the accessibility in general is not that important of a pull factor for national park visitors. Further research is needed to determine the relation between the accessibility and the visitations rates as indicated by the post density. Perhaps different accessibility datasets that have a higher resolution or are calculated specifically for the national parks at hand could prove useful.

The results on the point pattern and the Kernel density analyses in Sub-Saharan Africa show that the Flickr posts are fairly concentrated in the national parks of the region. Inspection of Serengeti and Nairobi national parks show that the most concentrated areas are located near the tourist attractions like the visitor centre in Serengeti or the elephant and rhino orphanage in Nairobi. Also, the roads and the paths seem to dictate the concentration of the Flickr posts. The comparatively lower clustering of posts in Nairobi national park might be caused by the extensive road network within the park. Further research could be done into the importance of roads as a driving factor of the point patterns by, for example, measuring how far the points are on average from the nearest road. Another plausible explanation for the concentration of the posts in parks like Serengeti may be that tourists visit the parks with dangerous wildlife as a part of a guided tour. Topography of the parks, which makes some sections unreachable, is also an important factor to consider in further research. Including accessibility data of different regions inside the park to the post concentration analysis could provide more answers for the clustering. Methods other that the ones used here (Ripley's K, KDE) for studying the patterns should be considered as well.

The linguistic analysis of the posts gives further insight on both the user base and what the visitors do at the parks. First, an overwhelming majority of the analysed posts (82 %) are made in English, which might indicate an overrepresentation of people from English speaking countries and in part explain the high post densities in North America and Europe. Though the analysis of Instagram posts from Finland reveals that the users post in English in significant amounts regardless of their home country (Hiippala et al., 2019). The country of origin of the posters and thus the number of tourists from each country could be estimated using a similar algorithm as Hiippala et al. use (2019, pp. 301). The simple word count analysis could be expanded in future research with a more complex topic modelling (see, e.g. Hiippala et al., 2019, pp. 303). This analysis could be used to learn more about the interest of the visitors even at a park-level.

In this article, we examined Flickr data in national parks worldwide. We first described the patterns the posts make at the regional and national scales. We learned that the link between the accessibility and the social media post density is not clear at the global or regional level. We then looked at the dispersion of the posts in the parks in Sub-Saharan Africa and found that clustering, as opposed to dispersing or forming random patterns, is extremely common. Finally, we described some features found in the linguistic content of the posts. We attempted to explain each result and offer ideas for future studies. In summary, we conclude that we achieved the study aims of both exploring the datasets and their possibilities in conservation science and offering methods for future research.

References

Africa Sun News. (2003). Africa national parks list. *Africa Sun News*. Retrieved 26.3.2020, available: http://www.africasunnews.com/national_parks.html

Alsaedi, N., Burnap, P. & Rana, O. (2016). *Temporal TF-IDF: A high performance approach for event summarization in Twitter*. IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, 2016. https://doi.org/10.1109/WI.2016.0087

Anselin, L. (2015). *Point pattern analysis: quadrat counts* [video file]. Retrieved 11.5.2020, available: https://www.youtube.com/watch?v=Ww95WKxUoZk

Balmford, A., Green, J. M. H., Anderson, M., Beresford, J., Huang, C., Naidoo, R., Walpole, M. & Manica, A. (2015). Walk on the wild side: estimating the global magnitude of visits to protected areas. *PLoS Biology 13*. https://doi.org/10.1371/journal.pbio.1002074

Bouton, S. N., Frederick, P. C., Dosualdo Rocha, C., Barbosa Dos Santos, A. T. & Bouton, T. C. (2009). Effects of tourist disturbance on wood stork nesting success and breeding behaviour in the Brazilian pantanal. *Waterbirds 28*, 487-497. https://doi.org/10.1675/1524-4695(2005)28[487:EOTDOW]2.0.CO;2

Buckley, R. (2009). Parks and tourism. *PLoS Biology 7(6)*. https://doi.org/10.1371/journal.pbio.1000143

Buckley, R. (2011). Tourism and environment. *Annual Review of Environment and Resources 36*, 397-416. https://doi.org/10.1146/annurev-environ-041210-132637

Buckley, R. C., Morrison, C. & Castley, J. G. (2016). Net effects of ecotourism on threatened species survival. *PLoS ONE 11*, 23-25. https://doi.org/10.1371/journal.pone.0147988

Cessford, G. & Muhar, A. (2003). Monitoring options for visitor numbers in national parks and natural areas. *Journal for Nature Conservation 11(4)*, 240-250. https://doi.org/10.1078/1617-1381-00055

Crush, J. S. (1980). National parks in Africa: a note on a problem of indigenization. *African Studies Review 23(3)*, 21-32. https://doi.org/10.2307/523669

Di Minin, E., MacMillan, D. C., Goodman, P. S., Slotow, R. & Moilanen, A. (2013). Conservation businesses and conservation planning in a biological diversity hotspot. *Conservation Biology 27*, 808-820. https://doi.org/10.1111/cobi.12048

Di Minin, E., Tenkanen, H. & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science 3*. https://doi.org/10.3389/fenvs.2015.00063

Eagles, P. F. J. & Wade, D. (2006). Tourism in Tanzania: Serengeti National Park. *Bois et forêts des tropiques 290(4)*, 73-80.

Encyclopaedia Britannica. (2020). National park. *Encyclopaedia Britannica.* Retrieved 13.3.2020, available: https://www.britannica.com/science/national-park

Elith, J., Graham, C. H., Anderson, R. P. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography 2,* 129-151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Frank, L., Bradley, M., Kavage, S., Chapman, J. & Lawton, T.K. (2008). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation 35 (1)*, 37–54. https://doi.org/10.1007/s11116-007-9136-6

García, R., Lopez, M., Pérez, S., Juan, C. & Raúl., P. (2015). *The kernel density estimation for the visualization of spatial patterns in urban studies*. 15th International Multidisciplinary Scientific GeoConference SGEM, Albena, Bulgaria, 2015 https://doi.org/10.5593/SGEM2015/B21/S8.111

Gillan, J. & Gonzalez, L. (2012). Ripley's K function and pair correlation function. *The Landscape Toolbox*. Retrieved 4.5.2020, available: https://wiki.landscapetoolbox.org/doku.php/spatial_analysis_methods:ripley_s_k_and_pair_correlation_function

Goodwin, H. (1996). In pursuit of ecotourism. *Biodiversity and Conservation 5*, 277-291. https://doi.org/10.1007/BF00051774

Gössling, S. (1999). Ecotourism: a means to safeguard biodiversity and ecosystem functions? *Ecological Economics 29(2)*, 303-320. https://doi.org/10.1016/S0921-8009(99)00012-9

Gössling, S. (2002). Global environmental consequences of tourism. *Global Environmental Change 12*, 283-302. https://doi.org/10.1016/S0959-3780(02)00044-4

Hart, T. & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. *Policing: An International Journal 37(2)*, 305-323. https://doi.org/10.1108/PIJPSM-04-2013-0039

Hausmann, A., Slotow, R., Fraser, I. & Di Minin, E. (2016). Ecotourism marketing alternative to charismatic megafauna can also support biodiversity conservation. *Animal Conservation 1*, 208-211. https://doi.org/10.1111/acv.12292

Hausmann, A., Toivonen, T., Heikinheimo, V., Tenkanen, H., Slotow, R. & Di Minin, E. (2017b). Social media reveal that charismatics species are not the main attractor of ecotourists to sub-Saharan protected areas. *Scientific Reports 7(1):763*. https://doi.org/10.1038/s41598-017-00858-6

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., Di Minin, E. (2017a). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters 11(1)*. https://doi.org/10.1111/conl.12343

Hiippala, T., Hausmann, A., Tenkanen, H., & Toivonen, T. (2019). Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities , 34(2)*, 290-309. https://doi.org/10.1093/llc/fqy049

Levin, N., Kark, S. & Crandall, D. (2015). Where have all the people gone? Enhancing global conservation using night lights and social media. *Ecological Applications 25*, 2153-2167. https://doi.org/10.1890/15-0113.1

Margules, C. R. & Pressey, R. L. (2000). Systematic conservation planning. *Nature 405,* 243-253. https://doi.org/10.1038/35012251

Mavoa, S., Witten, K., McCreanor, T. & O'Sullivan, D. (2012). GIS based destination accessibility via public transit and walking in Auckland, New Zealand. *Journal of Transport Geography 20 (1),* 15–22. https://doi.org/10.1016/j.jtrangeo.2011.10.001

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Mifflin Harcourt Publishing Company, New York, USA. Pp. 1-242.

Kaplan, A. & Haenlein. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons 53(1)*, 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Knight, A. T., Cowling, R. M. & Campbell, B. M. (2006). An operational model for implementing conservation action. *Conservation Biology 20*, 408-419. https://doi.org/10.1111/j.1523-1739.2006.00305.x

Krüger, O. (2005). The role of ecotourism in conservation: panacea or Pandora's box? *Biodiversity & Conservation 14*, 579-600. https://doi.org/10.1007/s10531-004-3917-4

Longley, P. A., Adnan, M. & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning 47(2)*, 465-484. https://doi.org/10.1068/a130122p

Pickering, C. M. & Hill, W. (2007). Impacts of recreation and tourism on plant biodiversity and vegetation in protected areas in Australia, *Journal of Environmental Management 85*, 791-800. https://doi.org/10.1016/j.jenvman.2006.11.021

Ranaweerage, E., Ranjeewa, A. D. G. & Sugimoto, K. (2015). Tourism-induced disturbance of wildlife in protected areas: a case study of free ranging elephants in Sri Lanka. *Global Ecological Conservation 4*, 625-631. https://doi.org/10.1016/j.gecco.2015.10.013

Richards, D. R. & Friess, D. A. (2015). A rapid indicator of cultural ecosystem service usage at a fine spatial scale: content analysis of social media photographs. *Ecological Indicators 53*, 187-195. https://doi.org/10.1016/j.ecolind.2015.01.034

Serengeti National Park. Serengeti National Park. *Serengeti National Park.* Retrieved 26.3.2020, available: https://www.serengetinationalpark.com/

Shuyo, N. (2010). *Language Detection Library for Java*. Retrieved 11.5.2020, available: https://github.com/shuyo/language-detection

Siegfried, W. R., Benn, G. A. & Gelderblom, C. M. (1998). Regional assessment and conservation implications of landscape characteristics of African national parks. *Biological Conservation 84(2)*, 131-140. https://doi.org/10.1016/S0006-3207(97)00110-9

Smith, R. J., Verissimo, D. & Macmillan, D. C. (2010). Marketing and conservation: how to lose friends and influence people. *In* Leader Williams, N., Adams, W. & Smith, R. (Eds.): *Trade-offs in conservation: deciding what to save,* pp. 215-232. Blackwells, Oxford, UK.

Su, S., Wan, C., Hu, Y. & Cai, Z. (2016). Characterizing geographical preferences of international tourists and the local influential factors in China using geo-tagged photos on social media. *Applied Geography 73*, 26-37. https://doi.org/10.1016/j.apgeog.2016.06.001

Steven, R., Pickering, C. & Castley, G. J. (2011). A review of the impacts of nature based recreation on birds. *Journal of Environmental Management 92*, 2287-2294. https://doi.org/10.1016/j.jenvman.2011.05.005

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L. & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports 7(1):17615.* https://doi.org/10.1038/s41598-017-18007-4

UNEP-WCMC (2019). *User manual for the World Database on Protected Areas and world database on other effective area-based conservation measures: 1.6*. UNEP-WCMC: Cambridge, UK. Retrieved 11.5.2020, available: http://wcmc.io/WDPA_Manual

UNEP-WCMC & IUCN (2020), *Protected planet: The World Database on Protected Areas (WDPA)* [On-line], version 01/2020, Cambridge, UK: UNEP-WCMC and IUCN. Retrieved 11.5.2020, available: www.protectedplanet.net

University of Helsinki. (2015). Social media data could contribute to conservation science. *ScienceDaily*. Retrieved 11.3.2020, available: https://www.sciencedaily.com/releases/2015/09/150915105007.htm

Watson, J. E. M., Dudley, N., Segan, D. B. & Hockings, M. (2014). The performance and potential of protected areas. *Nature 515,* 67-73. https://doi.org/10.1038/nature13947

Weiss, D. J. et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688), 333-336. https://doi.org/10.1038/nature25181

Willemen, L., Cottam, A. J., Drakou, E. G. & Burgess, N. D. (2015). Using social media to measure the contribution of red list species to the nature-based tourism potential of African protected areas. *PLoS ONE 10*. https://doi.org/10.1371/journal.pone.0129785

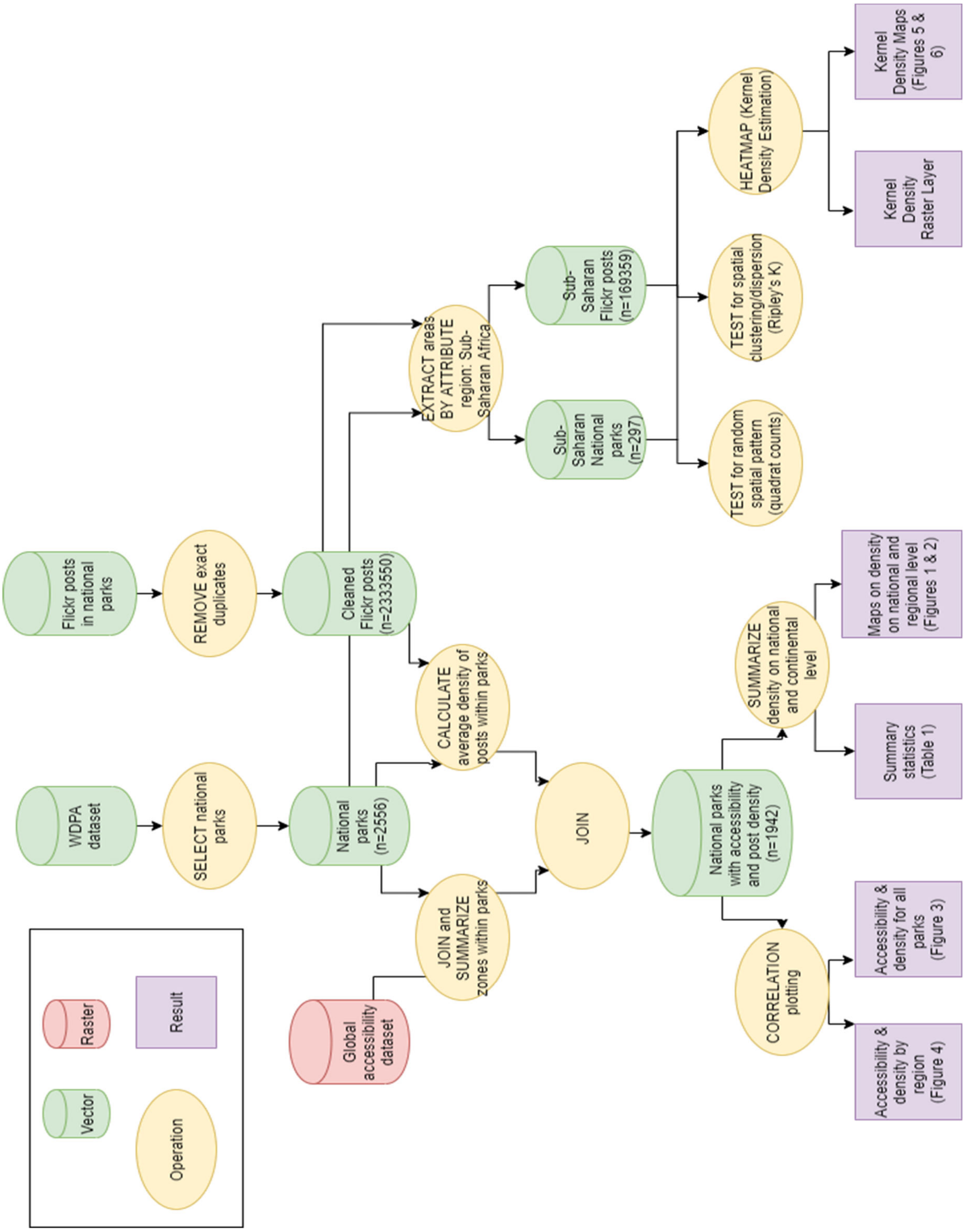
Woodroffe, R. & Ginsberg, J. R. (1998). Edge effects and the extinction of populations inside protected areas. *Science 280*, 2126-2128. https://doi.org/10.1126/science.280.5372.2126

World Tourism Organization. (2015). Towards measuring the economic value of wildlife watching tourism in Africa – briefing paper. *World Tourism Organization.*

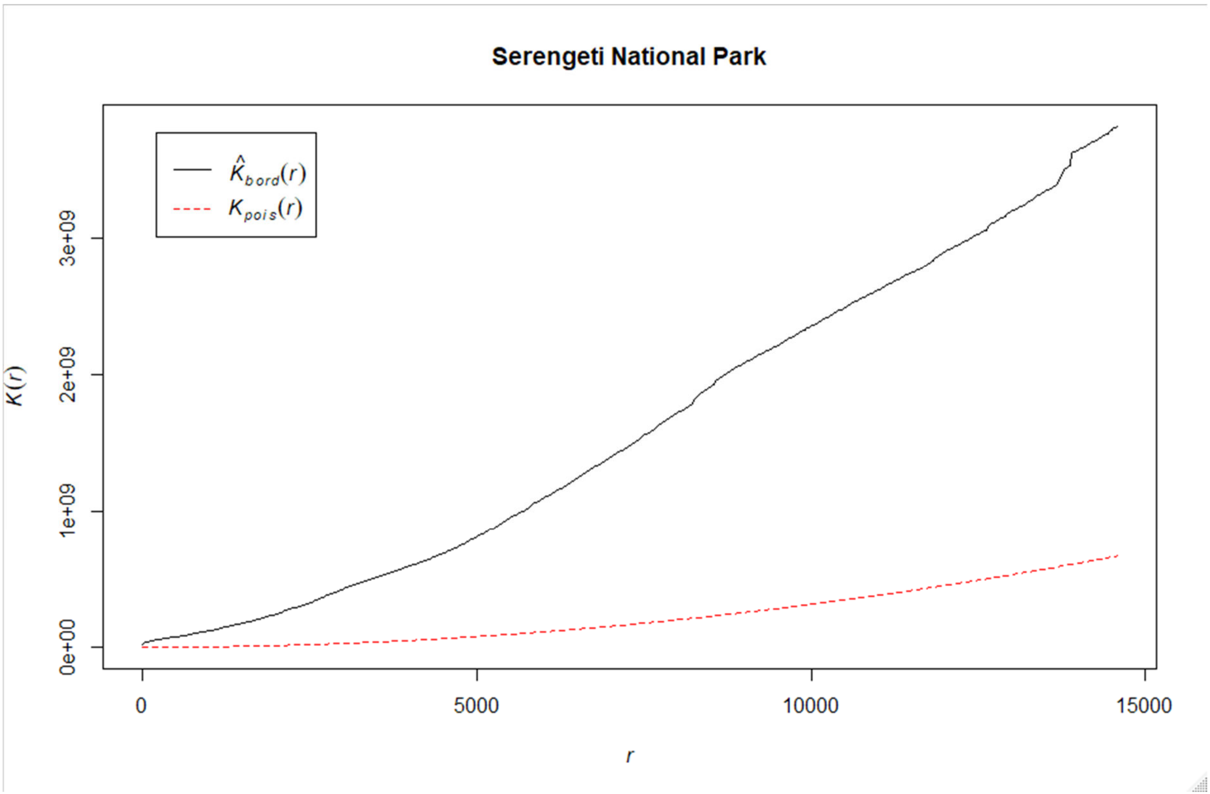
Hirvonen, H., Leppämäki, T., Rinne, J., Muukkonen, P. & Fink, C. (2020). Modifying and analyzing Flickr data for wildlife conservation. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 66–90. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Appendices

Appendix A: The spatial analysis workflow



Appendix B: The example plot of Ripley's K Function. Interpret it like this: if the black line is above the red dotted line (the result of a Poisson point process), it is clustered at that distance (X-axis). If below, it is dispersed and if roughly the same, the points are randomly located.



Examples and progress in geodata science presents the outcomes of a MSc level course *GEOG-G303 GIS project work*. The course was conducted in small working groups, each of which was assigned a separate project topic. The topics came from different research groups or teachers in the Department of Geosciences and Geography. These articles are written in close co-operation with researchers and teachers.

Department of Geosciences and Geography C19
ISSN-L 1798-7938
ISBN 978-951-51-4938-1 (PDF)
http://helda.helsinki.fi/

Helsinki 2020

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE